

Friedel Heinz Niemeier

**DIE BEDEUTUNG KOGNITIVER STRUKTUREN
FÜR DEN SCHULERFOLG**

**Methodenkritische Untersuchungen zu
Fragen der Forschungspraxis**

1982



Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41554825_0002

Psychologie

DIE BEDEUTUNG KOGNITIVER STRUKTUREN
FÜR DEN SCHULERFOLG

Methodenkritische Untersuchungen zu
Fragen der Forschungspraxis

Inaugural - Dissertation
zur Erlangung des Doktorgrades
der
P h i l o s o p h i s c h e n F a k u l t ä t
der
Westfälischen Wilhelms-Universität
zu
Münster (Westfalen)

vorgelegt von
Friedel Heinz Niemeier
aus Enniger
1982

Tag der mündlichen Prüfung: 8. Juli 1982

Dekan: Prof. Dr. M. E. Brockhoff

Referent: Prof. Dr. Lilly Kemmler

Korreferent: PD Dr. Ulf Plessen

Die Anregung zu dieser Arbeit entstand, als der Verfasser als wissenschaftlicher Mitarbeiter an einer Längsschnittstudie mitarbeitete, die 1960 von Frau Prof. Dr. Lilly Kemmler begonnen und bis zu deren Abschluß 1974 von ihr geleitet wurde.

Ich danke Frau Prof. Dr. Kemmler für die Anregung zu dieser Arbeit sowie für ihr sich über Jahre erstreckendes Interesse und die vielen Ratschläge und Hinweise, mit denen sie die Arbeit gefördert hat.

Inhaltsverzeichnis

T E I L I

1.0	EINLEITUNG UND FRAGESTELLUNGEN	1
2.0	ZUR METHODIK UNSERER UNTERSUCHUNGEN	13
3.0	DIE LÄNGSSCHNITTSTICHPROBE	17
3.1	Die Ausgangslage der Längsschnittstichprobe und deren sukzessive Schrumpfung	17
3.2	Die Längsschnittstichprobe zum Zeitpunkt der Untersuchung 1968	32
3.3	Die Klassifikationsmerkmale im einzelnen	36
4.0	DIE VERWENDETEN PSYCHOLOGISCHEN VERFAHREN	57
4.1	Ausgangsuntersuchung Ostern 1960 bzw. Ostern 1961	57
4.2	Untersuchung Frühjahr 1964	60
4.2.1	Intelligenztest (Testbatterie)	60
4.2.2	Fragebögen	66
4.3	Untersuchung Frühjahr 1968	69
4.3.1	Intelligenztest (Testbatterie)	70
4.3.2	Spezielle Fähigkeiten	73
4.3.3	Fragebögen	75
4.3.4	Schulleistungsvariablen	79
4.4	Untersuchung Frühjahr 1972 (nur Oberschüler)	80
5.0	DURCHFÜHRUNG DER UNTERSUCHUNGEN	82
5.1	Ausgangsuntersuchung Ostern 1960/1961	82
5.2	Untersuchung Frühjahr 1964	84
5.3	Untersuchung Frühjahr 1968	87
6.0	DIE FORMALE AUSWERTUNG DER TEST- UND FRAGEBOGENERHEBUNGEN	90
7.0	DIE INSTRUMENTELLEN EIGENSCHAFTEN DER VERFAHREN	93
7.1	Prüfung auf Normalverteilung	93
7.2	Exkurs: Bezogen auf die Eichung einiger bekannter Intelligenztests; HAWIE, HAWIK, IST, IST 70, LPS und PSB	108

7.3	Objektivität, Reliabilität, Validität	119
7.4	Objektivität	119
7.5	Reliabilität	120
7.5.1	Standardisierung der Tests	124
7.5.2	Standardmessfehler	125
7.5.3	Interkorrelationen der Testwerte	127
7.6	Validität	155
8.0	VERÄNDERUNGSMESSUNG KOGNITIVER MERKMALE AN KORRELIERTEN STICHPROBEN	167
8.1	Verteilungstatistische Differenzen der Meßwiederholungen bei verschiedenen Tests	174
8.2	Veränderungen nach Schülergruppen zusammengefaßt	192
8.3	Mittelwertdifferenzen und Regression zur Mitte	196
8.4	Extreme Aufsteiger und Absteiger im GSW	207
9.0	DIE DISKRIMINANZANALYSE ALS DIAGNOSE- UND PROGNOSEHILFE	218
9.1	Vorbemerkung	218
9.2	Diskriminanzanalyse der HS, RS, OS über 7 PMA 9-10	221
9.3	Diskriminanzanalyse der HS, RS, OS über 6 PMA 13-14	229
9.4	Diskriminanzanalyse der HS, RS, OS über GSW 1964	234
9.5	Weitere Diskriminanzanalysen	236
9.6	Diskriminanzanalyse der Schüler nach Schultart und Geschlecht, über 11 Leistungstests 1964	237
9.7	Diskriminanzanalyse der Schüler nach Schultart und Geschlecht, über 11 Leistungstests 1968	238
10.0	ZUM UMWELTEINFLUSS BEI DER INTELLI- GENZENTWICKLUNG	241
10.1	Schätzung der Umweltvarianz nach JENSEN	245
10.2	Analyse der Längsschnittdaten hin- sichtlich Umweltwirkungen auf die Intelligenzentwicklung	253
10.3	EXKURS	268

T E I L I I

DIE KONTROLLUNTERSUCHUNG

11.0	ÜBEREINSTIMMUNGSVALIDITÄT VON INTELLIGENZTESTS	274
11.1	Der Stellenwert der Kontrolluntersuchung	274
11.2	Die Kontrollstichprobe	276
11.3	Die verwendeten Tests, Durchführung und Auswertung der Kontrolluntersuchung	282
12.0	ERGEBNISSE DER KONTROLLUNTERSUCHUNG	287
12.1	Prüfung auf Normalverteilung	287
12.2	Validität der Tests	289
12.2.1	Interkorrelationen der PMA 13-14	289
12.2.2	Interkorrelationen der PSB	292
12.2.3	Korrelationen zwischen PMA 13-14 und PSB	293
12.3	Faktorenanalyse über 6 PMA 13-14, 9 PSB und FAT 7-8	294
12.4	Mittelwertunterschiede intra Kontrolluntersuchung und inter Kontroll- und Längsschnittuntersuchung	297
12.5	Regressionsanalysen über GL-PSB und GSW 13-14	301
12.6	Veränderungsmessung kognitiver Merkmale bei wiederholter Messung	302
12.7	Diskriminanzanalysen	306
13.0	ZUSAMMENFASSENDE INTERPRETATION DER ERGEBNISSE DER KONTROLLUNTERSUCHUNG	311
14.0	ZUSAMMENFASSENDE INTERPRETATION DER ERGEBNISSE DIESER ARBEIT UND AUSBLICK	316
	LITERATURVERZEICHNIS	347
	ANHANG	364

TEIL I

1.0

EINLEITUNG UND FRAGESTELLUNGEN

Die vorliegende Arbeit ist in zwei Teile gegliedert. Der erste Teil behandelt Aspekte einer Längsschnittstudie, welche die Begleitung der Entwicklung des schulischen Werdegangs einer Gruppe von Kindern vom Tag ihrer Einschulung (einschl. Zurückgestellte) bis zum Ende ihrer allgemeinbildenden Schulpflicht hatte. Hierzu wurden bereits umfangreiche und bedeutende Ergebnisse von KEMMLER veröffentlicht: "Erfolg und Versagen in der Grundschule" (KEMMLER 1967) und "Schulerfolg und Schulversagen" (KEMMLER 1976).

Bei der damaligen Arbeit am Forschungsprojekt fiel auf, daß wir uns des Öfteren vor "design"spezifischen Problemen sahen, von denen man in der Literatur nicht viel lesen kann, bzw. die als Störvarianz ohne bedeutenden Einfluß auf die Ergebnisse häufig übergangen werden. Wir denken dabei z.B. an die für alle Probanden gleiche Abfolge der Tests. Können damit nicht Versuchsfehler in Form von Positionseffekten eintreten? Oder die nicht minder bedeutsame Frage der Beschaffung und Zusammensetzung der Stichproben, bei Längsschnittstudien das besondere Problem der Wiederfindung der Ausgangsstichprobe: Liest man nicht häufig, eine Stichprobe erfülle zwar nicht ganz die Wünsche des jeweiligen

Autors, sei aber im übrigen repräsentativ und die Ergebnisse seien daher auch generalisierbar? Und dies trotz des oft gar nicht vermeidbaren Umstandes, daß Forscher häufig mit sog. Gelegenheitsstichproben vorliebnehmen müssen, die zumeist die Nachteile von Quotenstichproben in sich vereinen.

An dieser Stelle setzt unsere Arbeit an. Wir wollen uns mit dieser Situation nicht zufriedengeben, sondern den Versuch unternehmen, den Prozeß der Datengewinnung und -auswertung transparenter zu machen. Dies geschieht in der Absicht, eine Forderung, die METZGER 1952 für empirische Untersuchungen erhoben hat, einzulösen:

"Die Mitteilung auch der erstaunlichsten experimentellen Befunde ist wertlos, wenn nicht die Versuchsbedingungen so eingehend und anschaulich beschrieben werden, daß der Versuch vom Leser in der ursprünglichen Weise wiederholt und nachgeprüft werden kann. Zahlreiche widersprechende Ergebnisse vermeintlicher Nachprüfungen und daran sich knüpfende theoretische Auseinandersetzungen kamen nur dadurch zustande, daß infolge ungenügender Versuchsbeschreibungen der als Wiederholung beabsichtigte zweite Versuch in Wirklichkeit ein ganz anderer Versuch war" (METZGER 1952, S. 162, Hervorhebungen von W.M.).

Wir haben bei der Auflistung und Auswertung unserer Daten und bei den Literaturvergleichen den Eindruck gewonnen, daß in der Literatur viele Auseinandersetzungen mit empirischen Arbeiten von Kollegen - zustimmend oder kritisierend - vorgenommen wurden, ob-

gleich - mit METZGER zu reden - es sich in Wirklichkeit um ganz verschiedene Untersuchungen gehandelt hat, die aus den verschiedensten Gründen nicht vergleichbar waren.

Das Unbehagen mit der bisherigen Situation ist keineswegs neu, wie sonst hätten sich die Herausgeber von DIAGNOSTICA zu folgender Aussage genötigt sehen können:

"Das 'Ritual' gegenseitiger Bestätigungen hat in den vergangenen Jahren die Tendenz vieler Publikationen beherrscht. Wenn es zugunsten offener, kritisch-rationaler Diskussion gelockert (!) würde, könnte das den Fortschritt der wissenschaftlichen psychologischen Diagnostik sicher fördern" (JÄGER et al. 1979, S. 2).

Die Priorisierung der Ergebnisdiskussion rückt die vollständige Darstellung des "Designs" und der Methoden oft unverantwortlich weit in den Hintergrund. Diese Kritik richtet sich auch an die Herausgeber von Fachzeitschriften. Es wird offensichtlich häufig vergessen, daß man beinahe jeden beliebigen Effekt produzieren kann. Interessant wird es aber erst, wenn man weiß, wie er zustandegekommen ist. Die Schwierigkeiten beginnen bereits bei der Stichprobenbeschreibung. Wenn Leser einmal die Zusammensetzung einer Stichprobe nachzuvollziehen versuchen, kommt an Erkenntnis oft nicht mehr dabei heraus, als daß sich die

Stichprobe aus menschlichen Wesen und nicht aus Meerschweinchen zusammensetzt. Oder könnten Leser in der Regel mit den vorgefundenen Angaben den Versuch replizieren? Hier soll nicht dafür plädiert werden, die Freundlichkeit, die Anfang des Jahrhunderts Usus war, als Versuchspersonen noch namentlich in den Aufsätzen aufgeführt waren, wieder einzuführen.

Darauf, daß Angeführte mit Beispielen zu belegen, verzichten wir nur deshalb, weil es unfair wäre, nur einige Arbeiten zu zitieren. Der interessierte Leser wird u.E. keine Mühe haben, beliebig viele Beispiele zu finden.

Statt dessen sei es gestattet, einen Satz zu zitieren, den MELTON am Ende seiner Amtszeit als Herausgeber des "JOURNAL OF EXPERIMENTAL PSYCHOLOGY" im Dezember 1962 im JEP in einem Editorial geschrieben hat über die Auswahlkriterien für die Annahme von Artikeln. Eines der Kriterien war:

"... the assessment of an article involved a judgment with respect to the confidence to be placed in the findings - confidence that the results of the experiment would be repeatable under the conditions described. ... it reflects a belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounced."

Dieses Zitat macht nur Sinn, wenn das dort geforderte Kriterium praktisch häufig genug verletzt wurde. Inzwischen verfügt jedes Institut über mehr oder weniger große EDV-Anlagen, so daß dem Untersucher die - zugegeben - bei großen N's manchmal schweißtreibenden Kreuz- und Querausählungen abgenommen werden können. Es gibt also gar keinen sachlichen Grund mehr, sich mit den spärlichen Hinweisen auf die Repräsentativität von Stichproben zu beschränken (WEINERT ⁴1969, S. 13ff; LANGER/SCHULZ v. THUN 1974, S. 183; GADENNE 1976, S. 104; GNIECH 1976, S. 9ff; HELLER/ROSEMAN 1974, S. 69; JANSSEN 1979). Der Leser muß auf der Basis mitgeteilter Fakten (auch) zu einem anderen Urteil kommen können als der Autor. Die Grenze, wo blindes Vertrauen beginnen muß, wird häufig voreilig gezogen.

Offensichtlich bestehen bei vielen Untersuchern Hemmungen, sich "in die Karten gucken" zu lassen,

- sei es aus Angst, sich andernfalls ganz allgemein gegen Kritik nicht genug immunisieren zu können, und damit möglicher Abwertung der eigenen Forschung,
- sei es, etwas spezifischer, zugeben zu müssen, wie begrenzt die Reichweite der Resultate aufgrund der realisierten Forschungsbedingungen sei,
- oder sei es, eingestehen zu müssen, an sich banale Versuchsbedingungen verletzt oder mißachtet zu

haben.

Da wir diese Haltung nicht teilen, soll aus dem bisher Gesagten als erstes die Konsequenz gezogen werden, unsere Stichproben, insbesondere die Längsschnittstichprobe, so genau darzustellen, wie das die Daten zulassen. Dabei haben wir die allgemeine Hypothese, daß eine multivariate geschachtelte¹⁾ Beschreibung der Stichproben auf Probleme hinweist, die bei der üblichen ein- oder zweidimensionalen Darstellung von Stichproben häufig verdeckt bleiben, was die Möglichkeit in sich birgt, Effekte zu konstatieren, die aber u.U. nur Pseudoeffekte, also Stichprobenartefakte sind.

Zudem wollen wir das Unübliche tun und Unzulänglichkeiten bei der Durchführung dieser Untersuchung benennen und auf mögliche Folgen für unsere Ergebnisse hinweisen.

In diesem Zusammenhang ist auch die Frage von Bedeutung, wie die Merkmalsverteilung in den verwendeten Testvariablen beschaffen ist. Wo die Normalverteilung der Merkmale nicht gegeben ist, was waren

1) Geschachtelt meint die gleichzeitige Darstellung möglichst vieler Variablen.

die Gründe dafür, und könnten diese auch auf die bekannteren Intelligenztests zutreffen und mit welcher Bedeutung für diese Tests?

Großer Wert wird auf die Darstellung der korrelativen Zusammenhänge gelegt. Es soll gezeigt werden, daß die in Lehrbüchern und in der Literatur geläufige Behauptung, die Korrelation reduziere sich, wenn die Stichprobenvarianz eingeschränkt werde, so nicht haltbar ist. Vielmehr soll der besondere Stellenwert der Kovarianz in diesem Kontext herausgearbeitet werden.

Eine weitere hier behandelte Frage ist, ob Effekte der Veränderungsmessung (Wiederholungsmessung) notwendigerweise mit Hilfe des statistischen Phänomens der "Regression zur Mitte" gedeutet werden müssen oder ob sich psychologisch relevante Aspekte unterscheiden lassen.

Diese Frage wird mit Hilfe der Längsschnittmethode zu beantworten versucht, obgleich SCHAIE auf einem Symposium 1971 empfahl, die "traditional longitudinal method" ... "to lay to rest once and for all" (SCHAIE 1972, S. 3). Nach SCHAIE ist aufgrund einer Längsschnittanalyse nicht zu unterscheiden, ob sich eine Ver-

Änderung infolge der Veränderung des Individuums oder der Umwelt vollzogen hat. Auch wenn keine Veränderung stattgefunden habe, könne nicht entschieden werden, ob das fragliche Verhalten stabil geblieben sei oder ob Reifungsprozesse und Umwelteinflüsse einander aufgehoben hätten.

Für SCHAIE gibt es fünf gewichtige technische Gründe gegen die Längsschnittmethode (1972, S. 7-13):

- "(1) experimental mortality,
- (2) misleading regression effects,
- (3) design complication in handling the effects of prior measurements,
- (4) lowered efficiency of repeated measurement design, and
- (5) changes in factor structure over time in multivariate studies."

SCHAIE belegt bzw. verbindet seine Ausführungen allerdings häufig mit Beispielen aus dem Erwachsenenalter.

So wichtig die aufgezeigten Schwächen gegen Längsschnittuntersuchungen auch sein mögen, sie haben jedenfalls bisher nicht zur "Grablegung" der Längsschnittmethode unter Anteilnahme der Forschergemeinde geführt, da SCHAIE bisher nicht hat nachweisen können, daß mit Hilfe seiner seit 1965 vertretenen

Sequenz-Modelle alle Entwicklungsprobleme lösbar wären (vgl. HINDLEY 1972, S. 38ff; McCALL 1977, S. 341; THOMAE 1979, S. 38-40). Im Gegensatz zu SCHAIK wird hier die Auffassung vertreten, daß Längsschnittuntersuchungen sehr wohl ihre Daseinsberechtigung haben. Sie eignen sich z.B. sehr gut zu Forschungen, die der Schullaufbahn der Probanden gelten. In den letzten Jahren wird zunehmend nach Wegen und Methoden gesucht, die Schullaufbahnberatung zu verbessern (vgl. HELLER 1976; ROSEMAN 1977). Sie dient dem Zweck, den Schülern mehr Informationen für eine möglichst problemfreie, erfolgreiche - dabei der vollen Entfaltung der Persönlichkeit dienend - individuelle Wahl des Weges durch das allgemeinbildende Schulwesen zu vermitteln. Wir gehen daher hier der Frage nach, ob die multiple Diskriminanzanalyse als leistungsfähiges Hilfsmittel zur Diagnose und Prognose bei der Schullaufbahnberatung empfohlen werden könne.

In einem letzten Kapitel des ersten Teils dieser Arbeit gehen wir einer Frage nach, die auf den ersten Blick etwas abseits zu liegen scheint. Es soll untersucht werden, ob mit Hilfe von Längsschnittdaten Umweltwirkungen auf die Intelligenzentwicklung durch die Bildung von Pseudozwillingen nachgewiesen werden können.

Fragen die der Intelligenzentwicklung gelten, werden oft mit sehr prinzipieller Kritik bedacht. Sie richtet sich häufig generell gegen Intelligenzkonzepte und Testverfahren zu deren Messung. Die allgemeine Kritik ist, daß die Psychologen nicht wüßten, was denn Intelligenz überhaupt sei, sie aber dennoch messen zu können behaupteten und dies auch fleißig praktizierten.

Man könnte sich nun hinter der Mathematik verstecken, einer Disziplin, von der gemeinhin nicht angenommen wird, sie operiere mit nicht definierten Begriffen. In einem Kapitel über Definitionen heißt es jedoch bei MESCHKOWSKI (1973, S. 60), nachdem er über die Unzulänglichkeiten bestimmter Definitionen referiert hat:

"Die moderne Mathematik hat daraus die Konsequenz gezogen, daß man auf die Definition der geometrischen Grundbegriffe (Punkt, Gerade, Ebene) verzichten müsse."

Dennoch können die Mathematiker trefflich mit "Punkt" und "Gerade" operieren. Mit ähnlichen Problemen schlagen sich vermutlich alle Wissenschaften herum. Das Fehlen einer allseits befriedigenden Definition kann demnach kein hinreichendes Kriterium gegen die Verwendung eines Begriffes sein. Man würde den Erkenntnisprozeß letztlich paralysieren.

Und da es aber falsche Definitionen nicht gibt,

sondern nur zweckmäßige und unzweckmäßige, ist die Frage zu stellen, ob es zweckmäßig sei, "die Intelligenz" zu definieren, oder ob es nicht vorteilhafter sei zu versuchen, jeweils die einzelnen Fähigkeiten zu definieren, die im Rahmen einer bestimmten Intelligenztheorie erfaßt werden sollen? Zudem sind die kontroversen Publikationen um den Intelligenzbegriff wohl nur wegen dessen Wertlosigkeit so harsch.

Wir teilen bei aller Unvollkommenheit und Vorläufigkeit unseres Instrumentariums die Feststellung MEILI's auf dem Diagnostik-Symposium 1975:

"Ich glaube nicht, dass ein Psychologe, Berater usw. verantwortungsvoll arbeitet, wenn er nicht so viel wie möglich solche Instrumente (Anm.: standardisierte Tests) verwendet. Dass das nicht genügt, dass er darüber hinausgehen muß, ist selbstverständlich, ..." (MEILI 1978, S. 203f).

Bei allen berechtigten Vorbehalten wegen der unleugbaren Unzulänglichkeiten bleibt die Frage, wie die sich in der Praxis des Schulalltags stellenden Probleme auf andere Weise wirkungsvoller gelöst werden könnten.

In diesem Kontext spielt auch die Frage der Umweltbedingungen auf die Intelligenz eine Rolle. In der

Anlage-Umwelt-Kontoverse der Intelligenzentwicklung wird versucht, die Umweltwirkungen u.a. durch die Varianz von nicht verwandten Probandengruppen zu schätzen (JENSEN 1969). Die vorliegende Längsschnittstudie bietet die Möglichkeit, hinsichtlich soziodemographischer Merkmale relativ homogene Untergruppen und daraus Pseudozwillingspaare zu bilden und so auch im Längsschnittaspekt die Umweltwirkungen zu schätzen.

Im Teil II dieser Arbeit werden die Ergebnisse einer Querschnittuntersuchung wiedergegeben. Ziel der Querschnittuntersuchung ist, die Übereinstimmungsvalidität der in der Längsschnittstudie verwendeten PMA 13-14 mit einem geeichten Test, dem PSB von HORN zu ermitteln. Gleichzeitig wird hier das Problem der Veränderungsmessung nochmals aufgenommen. Schließlich wird die Diskriminanzanalyse als Hilfsmittel der Schullaufbahnberatung auch auf die Querschnittstichprobe angewandt.

ZUR METHODIK UNSERER UNTERSUCHUNGEN

Im Kapitel "Einleitung und Fragestellungen" haben wir dargestellt, welche Fragen Gegenstand dieser Arbeit sein sollen. Dabei ist dem Leser sicher nicht entgangen, daß zumindest einige Fragestellungen etwas allgemein gehalten sind. Dazu bedarf es einiger Erläuterungen. Wie bereits ausgeführt, verdanken wir unsere Daten einer Längsschnittstudie zur Problematik "Schulerfolg und Schulversagen", und einer Kontrolluntersuchung, deren Datenerhebungen zu Beginn unserer Untersuchung abgeschlossen waren. Auf der folgenden Seite geben wir eine Zusammenstellung der Untersuchungstermine, der jeweils untersuchten Schülerzahlen sowie Altersangaben der Probanden zur Zeit der Untersuchungen wieder.

Die Längsschnittstichprobe, die verwendeten psychologischen Verfahren, die Durchführung und Auswertung der Untersuchungen, sowie die instrumentellen Eigenschaften der Verfahren werden in getrennten Kapiteln beschrieben und analysiert. Dies gilt auch für die Kontrolluntersuchung.

In der Einleitung ihres ersten Buches zur Längsschnittstudie schreibt KEMMLER:

"In der nachfolgenden Untersuchung kann es sich nicht darum handeln, eine einzelne Hypothese aufzustellen, die dann durch die Befunde entweder bestätigt oder verworfen wird. Wir werden es mit einem Geflecht von Einzelbefunden, mehr oder minder vagen Vermutungen, mit pädagogischen Theorien und klinischen Verallgemeinerungen zu tun haben. Wenn man so will, handelt es sich bei diesem Unternehmen um einen großangelegten Erkundungsversuch" (KEMMLER 1967, S. 13).

Zusammenstellung der Untersuchungstermine, der jeweils untersuchten Schülerzahl, sowie Altersangaben der Probanden zur Zeit der Untersuchungen

Längsschnittuntersuchung		
Unter- suchung	Gesamtzahl der Probanden N = 341 (20 Schüler konnten nicht erfaßt werden)	
1960	altersgemäß Eingeschulte N = 250	Alter der Probanden 5;10 - 6;9 Jahre Stichtag: 30.4.1960
1961	Zurückgestellte N = 71	6;10 - 7;9 Jahre Stichtag: 30.4.1961
1964	N = 251	9;7 - 10;6 Jahre Stichtag: 31.1.1964
1968	N = 224	13;6 - 14;5 Jahre Stichtag: 15.12.1968
1972	N = 46 (nur Oberschüler)	17;9 - 18;10 Stichtag: nicht festgelegt
=====		
Kontrolluntersuchung		
1970	N = 116	13;6 - 14;5 Jahre Stichtag: 15.10.1970

Damit ist auch der Rahmen abgesteckt, innerhalb dessen sich unsere Fragestellungen bewegen können. Es wäre u.E. wissenschaftsethisch nicht vertretbar, wenn wir etwa Fragen im Sinne strenger Hypothesentestung formuliert hätten. Die Untersuchungen hätten dazu gezielt angelegt sein müssen.

Zwar streben Forscher danach, eine Theorie bzw. Hypothesen zu haben, die sie testen können, weil letztlich

in der Überprüfung einer konkreten Theorie der eigentliche Erkenntniszuwachs gesehen wird. Aber in der pädagogischen Psychologie, der unsere Fragestellungen zuzuordnen sind, ist es bisher noch nicht gelungen, z.B. eine testbare Theorie des "Schülers im allgemeinbildenden Schulsystem" zu formulieren. Damit meinen wir eine integrierte Theorie, die alle Bedingungen für Schulerfolg umfaßt und es erlaubt, im Zuge der Schullaufbahnberatung Schulerfolg in dem Sinne zu prognostizieren, daß der einzelne Schüler den größtmöglichen persönlichen Nutzen von solch einer Beratung erwarten könnte und die volle, freie Entfaltung seiner Persönlichkeit in diesem Bereich gewährleistet wäre.

Wichtige Teilbereiche für solch eine Theorie können bereits als gut erforscht gelten. So hat z.B. KEMMLER in dieser Längsschnittstudie die große Bedeutung der Verbalität der Schüler für eine möglichst hochwertige Schulbildung in unserem Schulsystem nachweisen können. Andere Bereiche weisen demgegenüber noch mehr oder weniger große Defizite auf. So ist z.B. nicht hinreichend erforscht, ob und wie sich kognitive Fähigkeiten von 10jährigen Kindern verändern und welche Rolle dabei den Schulen zukommt. Auch über die Anzahl unterscheidbarer kognitiver Fähigkeiten besteht noch kein Konsens. Andererseits sind für den Praxisbereich Methoden entwickelt oder aus anderen Gebieten übernommen worden, von denen angenommen wird, daß sie helfen könnten, den

Schullaufbahnberatungs- und Entscheidungsfindungsprozeß zu optimieren. Bevor solche Methoden jedoch eingesetzt werden können, müssen sie sich bewährt haben. Zum Nachweis der Brauchbarkeit sind oft längerfristige Untersuchungen notwendig.

Neben dem Wunsch, einige formale Aspekte der Forschungspraxis, die Einfluß auf die inhaltlichen Ergebnisse von Forschung haben, einer kritischen Analyse zu unterziehen, haben wir uns die Frage gestellt, ob nicht die seltene Chance genutzt werden könne, nämlich den Vorteil der Vorlage von Meßwiederholungsdaten und schulbegleitenden Untersuchungen, den KEMMLER's Längsschnittstudie bietet, zur Untersuchung einiger Fragen nutzbar zu machen, wozu sonst neue Datenerhebungen notwendig wären. Diese Frage konnten wir bejahen. Unsere Fragen haben - entsprechend der Gesamtkonzeption der Längsschnittstudie - ebenfalls explorativen Charakter. Zur Herausarbeitung unserer Ergebnisse werden sowohl Verfahren der deskriptiven Statistik wie auch der Inferenzstatistik verwendet. Dabei haben wir die Hoffnung, daß sich auf Grund unserer Ergebnisse einige Fragen vielleicht im Hinblick auf zukünftige Hypothesentestung konkretisieren lassen.

3.0 DIE LÄNGSSCHNITTSTICHPROBE

3.1 Die Ausgangslage der Längsschnittstichprobe und deren sukzessive Schrumpfung =====

Wir werden hier eine multivariate - geschachtelte - Darstellung der Längsschnittstichprobe vornehmen. Für die Zeit von 1960 - 1964 verweisen wir auch auf die Arbeiten von ZINGELER (1965) und KEMMLER (1967).

Zur Bewertung und Einordnung der Ergebnisse muß man bzgl. der Ausgangssituation u. a. wissen: "Zweck der Untersuchung war es, eine größere Anzahl von Schülern über vier Schuljahre, also bis zum Übertritt in die Oberstufe der Volksschule (Hauptschule) oder bis zum Übergang auf eine höhere Schule, in ihrer schulischen, intellektuellen und seelisch-sozialen Entwicklung zu verfolgen" (KEMMLER 1967, S. 30). Dieses Zitat weist aus, daß die Untersuchung ursprünglich nicht über den nun doch vollzogenen Zeitraum durchgeführt werden sollte, eine Ausgangslage, die jedoch (verständlicherweise) nicht unerheblichen Einfluß auf die Planung und Durchführung der einzelnen Datenerhebungen hatte. Wie noch zu zeigen sein wird, wurde auf die Erhebung von für die damalige Planung entbehrlicher, für die Längsschnittstudie in ihrer jetzigen Form jedoch relevanter psychologischer und soziologischer Daten verzichtet, was zur Folge hatte, daß wir manchen Fragen - auch solchen, die sich erst bei der Gesamtanalyse stellten - nicht in der gewünschten

Weise nachgehen konnten. Details hierzu werden an passender Stelle referiert.

Zur Frage der Herkunft der Stichprobe zitieren wir noch einmal KEMMLER (1967, S. 30): "In Zusammenarbeit mit dem Schulamt der Stadt Münster wurden für die Untersuchung 5 Volksschulen in den Randbezirken der Stadt ausgewählt. Damit sollte ein zu großer Anteil von Kindern aus gehobenem Milieu, wie es für Münster typisch ist (höhere Beamte, Akademiker), vermieden werden.

Im Frühjahr 1960 hatten im Bezirk dieser 5 Schulen 341 Kinder das schulpflichtige Alter erreicht: d. h. am 30. 04. 1960 - dem Stichtag der ersten Untersuchung - hatten alle Kinder ein vollendetes Alter von 5;10 - 6;9 Jahren."

Infolge Zurückstellung vom Schulbesuch, entweder vor oder kurz nach der Einschulung, ergaben sich zwei Primärstichproben, die wir mit KEMMLER (1967) "altersgemäß Eingeschulte" und "Zurückgestellte" bezeichnen. Um bezüglich der schulischen Unterrichtsdauer (in Schuljahren) möglichst homogene Gruppen zu erhalten, haben wir für diese Analyse - im Gegensatz zu KEMMLER (1967) - die (kurz) nach der Einschulung noch zurückgestellten Schüler zu den Zurückgestellten gezählt.

20 von diesen 341 Schülern haben aus nicht vollständig bekannten Gründen nie an einer der Untersuchungen teil-

genommen. Wir haben somit einen Totalverlust von ca. 6%, von dem wir kaum Informationen besitzen. Dieser Verlust geht beinahe vollständig zu Lasten der Zurückgestellten. Ob es sich hierbei über die von uns - nachfolgend - auf Zufälligkeit der Stichprobenschrumpfung geprüften Klassifikationsmerkmale hinaus um eine systematische Auslese handelt, ist nachträglich nicht mehr rekonstruierbar. Allerdings sei vermerkt, daß es sich bei dieser Verlustgruppe zum Teil aus zweimal zurückgestellten und - mindestens aus damaliger Sicht - vermindert bildungsfähigen Kindern handelt (KEMMLER 1967, S. 31).

Tabelle 1 zeigt nun die Zusammensetzung der Längsschnittstichprobe zu den 3 Untersuchungsterminen, jeweils aufgeschlüsselt nach Einschulungstermin, Geschlecht und Alter. Dargestellt sind die absoluten Frequenzen und die relativen (Prozente), bezogen auf das jeweilige Gesamt-N des betreffenden Untersuchungszeitpunktes. Ein Vergleich analoger Prozentwerte der verschiedenen Untersuchungszeitpunkte zeigt, daß die Schrumpfung der Stichprobe hinsichtlich der dargestellten Kriterien - erwartungsgemäß - zufällig ist. Auffällig ist jedoch der große Unterschied in der Anzahl der jüngeren Jungen gegenüber derjenigen der älteren Mädchen (Spalte $\sum f$). Die Unterschiede sind zu allen Untersuchungszeitpunkten überzufällig.⁺⁾ Wie ersichtlich, spielen die zurückgestellten Schüler/innen eine besondere Rolle, worauf bei der weiteren Analyse wieder Bezug zu nehmen sein wird.

⁺⁾ Hier und im folgenden wurde die Analyse der Stichprobenzusammensetzung immer mit dem Chi-Quadratstest überprüft. Dabei betrachten wir Unterschiede mit einer Irrtumswahrscheinlichkeit $p < 5\%$ als überzufällig, bedeutsam oder signifikant.

Tabelle 1: Zusammensetzung der Längsschnittstichprobe zu den verschiedenen Untersuchungszeitpunkten

1 := jüngeres Halbjahr, 2 := älteres Halbjahr
m := männlich, w := weiblich

Unter- suchung	Alter Genus	Altersgemäß Eingeschulte		Zurückgestellte		Gesamt		
		f	f%	f	f%	Σf	f%	
1960/61	1	m	67	20,8	37	11,5	104	32,3
		w	50	15,6	25	7,8	75	23,4
	2	m	70	21,8	7	2,2	77	24,0
		w	63	19,6	2	0,6	65	20,2
	Summe		250 ⁺	77,8	71 ⁺	22,1	321	99,9
	1964	1	m	49	19,5	31	12,3	80
w			40	15,9	20	8,0	60	23,9
2		m	54	21,5	5	2,0	59	23,5
		w	51	20,3	1	0,4	52	20,7
Summe		194	77,2	57	22,7	251	99,9	
1968		1	m	45	20,1	25	11,1	70
	w		37	16,5	19	8,5	56	25,0
	2	m	48	21,4	5	2,2	53	23,6
		w	44	19,6	1	0,5	45	20,1
	Summe		174	77,6	50	22,3	224	99,9

⁺) Abweichungen zu KEMMLER 1967, S. 30f, ergeben sich dadurch, daß in dieser Tabelle auch die nachträglich zurückgestellten Schüler zu den Zurückgestellten gezählt werden (s. o.).

Aus der Anfangsstichprobe sind insgesamt - bezogen auf die jeweiligen Ausgangsgruppen der Geschlechter - 32% Jungen und 28% Mädchen ausgeschieden.

An dieser Stelle muß jedoch darauf hingewiesen werden, daß wir den vermutlich wichtigsten Indikator für die Art der Stichprobenschrumpfung - den Beruf des Vaters der Schüler/innen als Maß für die regionale Mobilität - nicht berücksichtigen konnten, da die Berufe erst 1968 erfragt wurden. Es ist bekannt, daß der Ober- und Mittelschicht eine größere regionale Mobilität eigen ist (BOLTE 1959). Mithin kann trotz und wegen obiger Ausführungen eine spezifische Schrumpfung nicht ausgeschlossen werden. Dagegen spräche allenfalls die besondere Attraktivität der Stadt Münster für die Untersuchungsregion.

Tabelle 2 soll nun zeigen, wie sich unsere Untersuchungsgruppe - bezogen auf die 5 Grundschulen - von 1968 auf die Ausgangssituation 1960/61 zurückblickend - zusammenfügt. Dabei haben wir auch die Anzahl der aus der Stichprobe ausgeschiedenen Schüler/innen grundschulspezifisch in die Tabelle aufgenommen. Ausgeschieden heißt dabei in aller Regel; aus der Untersuchungsregion verzogen bzw. für uns nicht mehr erreichbar.

Tabelle 2: Retrospektive Zusammensetzung der Längsschnitt-
stichprobe vom Zeitpunkt der Untersuchung 1968
(vor dem Versetzungstermin)

Grund- schule	Alter	Genus	Altersgemäß Eingesch. (1960 eingeschult)					Zurückgestellte (1961 eingeschult)						
			SS	HS	RS	OS	ausg	SS	HS	RS	OS	ausg		
1	Jü	m		8	1	7	2	2	1				21	67
	w		2	4	5	1	1	2		1		16		
	Äl	m		4	3	5	1 1	1				13		
	w		8	7		2 1						17		
2	Jü	m		1	2	2	5		3	1	1		14	27
	w		1	1	2						1	4		
	Äl	m				2	2					4		
	w	1	2		1	1						5		
3	Jü	m		2	4	4	5		1 1	1	4	6	27	85
	w		2	3	3	4		1 4	1	2	2	21		
	Äl	m		2	2	1 1	2 6					21		
	w		3	1 2	6	2 5						16		
4	Jü	m		3		3	2	1	4		2	2	17	66
	w		5	2	2	1 2	1	5			1	18		
	Äl	m		1 4	1	4	2	2	1			14		
	w		6 1	1 4	2	1 5						17		
5	Jü	m		1 4		5	1 3	2	3			4	21	70
	w		2	3	3	4 1		2			1	15		
	Äl	m		6 1	1 7	1	7		1			2	24	
	w		2	3	1	2	1				1	10		
			1	67	48	69	60	11	27	2	10	20		
			245					70					315	

Erklärungen s. Folgeseite

Erklärungen zu Tabelle 2:

Ziffern 1 - 5 := Volks-(Grund-)schulcode

JÜ := Jüngeres Halbjahr

Äl := Älteres Halbjahr

m := männlich

w := weiblich

SS := Sonderschule / -schüler

HS := Hauptschule / -schüler

RS := Realschule / -schüler

OS := Oberschule / -schüler, (Gymnasien)

ausg := aus der Untersuchung ausgeschieden, d. h. für
die Untersucher nicht mehr erreichbar.

Die in einigen Zellen tiefer gesetzten Häufigkeitsangaben bedeuten:

- nach links tiefer := aus OS in die betreffende Schule

- nach rechts tiefer := aus RS in die betreffende Schule

In diesem Kapitel werden eine Reihe von Chi-Quadrat-Berechnungen durchgeführt. Zur Information des Lesers werden die Häufigkeitstabellen im Text wiedergegeben, die Tabellen aber bewußt nicht als Tabellen deklariert.

Es können an Hand der Tabelle 2 folgende Fragen beantwortet werden:

Unterscheiden sich die Substichproben aus den 5 Grundschulen hinsichtlich

- a) des Einschulungstermins,
- b) des Geschlechts,
- c) des Alters (in Halbjahren),
- d) des Anteils an den später besuchten 4 Schularten,
- e) der Quoten der ausgeschiedenen Pbn?

Bedeutsame Verteilungsunterschiede in den Variablen Geschlecht und Alter lassen sich nicht nachweisen. Anders ist dagegen die Verteilung der Schüler/innen auf die 4 Schularten: Bei Zusammenfassung der SS und HS (aus formalen Gründen) ergibt die Chi-Quadratprobe einer $5 \times 3 = 15$ Feldertafel eine sehr signifikante Gleichverteilungsabweichung. Sie wird im wesentlichen dadurch verursacht, daß viel mehr Schüler der Grundschule 3 eine weiterführende Schule besucht haben (43 von 57 der nach der Grundschulzeit noch erreichten), während aus der Grundschule 4 erwartungswidrig wenige Schüler in diesen Vorzug gekommen sind (20 von 52). Andererseits gehen aus den Grundschulen 4 und 1 je 4 Schüler (je 1/3 aller SS) später zur Sonderschule.

Wir konnten von 235 Schülern in Erfahrung bringen, welche Schulbildung sie nach der Grundschulzeit begonnen haben: 12 SS (= 5% von 235), 94 HS (= 40%), 50 RS (= 21%), 79 OS (= 34%). Bezogen auf 315 lauten die entsprechenden Prozentzahlen 4, 30, 16, 25; insgesamt haben somit 55% (von 235) bzw. 41% (von 315) der (zunächst) in oder um Münster ver-

bliebenen Schüler eine RS oder QS Schulbildung erstrebt. Die auf N=315 bezogenen Prozente sind Mindestwerte, da angenommen werden kann, daß von den "Ausgeschiedenen" auch noch Schüler zur RS oder QS gegangen sind.

Die Verteilung des Stichprobenschwundes zeigt ebenfalls ein nicht erwartetes Ergebnis:

Grundschule	1	2	3	4	5	Summe
bis einschl. 1968 unter- suchte Schüler	59	18	53	50	44	224
vor 1968 aus- geschiedene Schüler	8	9	32	16	26	91
Summe	67	27	85	66	70	315 ⁺⁾

⁺⁾ (warum hier nur 315 Vpn eingegangen sind, wird unten erläutert.)

Die Chi-Quadratprobe wird sehr signifikant. Das bedeutet Nicht-Zufälligkeit des Schwundes in den einzelnen Grundschulen. Sie geht hauptsächlich zu "Lasten" der Schule 1 (zu geringer Schwund) und der Schulen 3 und 5 (zu großer Schwund). Für mögliche Ursachen sollen im Zusammenhang mit der Darstellung der Väterberufe der Schüler/innen einige Antworten gegeben werden.

Aber schon die Einschulung bzw. die Zurückstellung bei der Einschulung wurde nicht von allen Grundschulen in gleicher Weise gehandhabt. Obgleich hinsichtlich des Geschlechts und des Alters keine interschulisch bedeut-

samen Unterschiede bestehen, ist die Chi-Quadratprobe der folgenden $5 \times 2 = 10$ Feldertafel hoch signifikant, d. h. aus den Grundschulen 1 und 2 wurden erwartungswidrig wenige und aus den Schulen 3, 4 und 5 erwartungswidrig viele Schüler bei der Einschulung um ein Jahr vom Schulbesuch zurückgestellt.

Grundschule	Einschulung		Summe
	1960	1961	
1	66	1	67
2	22	5	27
3	64	21	85
4	47	19	66
5	53	17	70
Summe	252	63	315

Um auch über psychologische Variablen Hinweise zur Stichprobenschrumpfung zu gewinnen, haben wir zusätzliche Berechnungen durchgeführt (Beschreibung der psychologischen Verfahren s. S.).

Verglichen wurde nun immer die Schrumpfungsquote (N_2) zwischen 2 bzw. 3 Untersuchungsterminen mit den jeweils noch erreichten Pbn (N_1). Als Testvariablen dienten dabei logischerweise immer die Verfahren, die von allen (miteinander verglichenen) Pbn bearbeitet worden waren. Für 1960/61 - 1964 waren dies die Gesamtrohwerte des

Schulreifetests von KERN (1957) und des Münchener-Schulreifetests von WALTER (o.J.), Standardwerte wurden 1960/61 nicht berechnet); für 1964 - 1968, außer den bereits genannten, die Unterstandardwerte der PMA 9 - 10 (Faktoren: Verbal, Reasoning, Lesefrei) sowie der Gesamtstandardwert; für 1960/61 - 1968 wiederum nur die beiden genannten Schulreifetests.

Tabelle 3 zeigt die Zusammenstellung der Ergebnisse, die - bezogen auf die Mittelwerte - alle nicht signifikant; d. h. bezogen auf diese psychologischen Variablen ist die jeweils verbliebene Stichprobe bei den Folgeuntersuchungen vergleichbar mit der Ausgangsstichprobe.

Wir haben bereits dargelegt, daß insgesamt $N = 321$ Probanden (Pbn) an der Untersuchung teilgenommen haben. Die Fülle der Daten war nur noch mittels elektronischer Rechenanlage analysierbar, so daß die Daten auf ein Magnetband gespeichert wurden. Dabei fiel auf, daß für 6 Pbn schon von der ersten Untersuchung 1960/61 die Daten unvollständig waren, so daß diese Pbn bei den meisten Vergleichen hätten eliminiert werden müssen. Dies veranlaßte uns dazu, die Stichprobe um diese 6 Vpn zu reduzieren, d.h. auf unserem Magnetband befinden sich Informationen von $N = 315$ Probanden.

KEMMLER (1976, S. 21) geht von $N=314$ aus. Die fehlende Vp ist unmittelbar nach Beginn der 1. Untersuchung aus der Gesamtstichprobe ausgeschieden.

Tabelle 3: Prüfung des Stichprobenschwundes auf Zufälligkeit
im Untersuchungszeitraum 1960/61 - 1968

Schumpfung von	Variablen	Vergleichs- gruppen	Mittelwert		Streuung	
			N ₁	N ₂	N ₁	N ₂
1960/61 bis 1964	RW KERN-Ges.	N ₁ = 251	1,83	1,78	1,76	1,58
	RW Mün.-Ges.	N ₂ = 64	37,43	35,91	12,34	12,80
1964 bis 1968	RW KERN-Ges.	N ₁ = 224	1,87	1,56	1,80	1,42
	RW Mün.-Ges.	N ₂ = 27	37,65	35,67	12,42	11,76
	PMA 9 - 10 USW-Verbal		49,61	48,70	10,28	10,83
	PMA 9 - 10 USW-Reasoning		50,19	49,26	10,45	9,59
	PMA 9 - 10 USW-Lesefrei		50,52	48,37	10,01	12,65 ⁺)
	PMA 9 - 10 GSW		51,17	49,56	10,60	13,06
1960/61 bis 1968	RW KERN-Ges.	N ₁ = 224	1,87	1,71	1,80	1,53 ⁺)
	RW Mün.-Ges.	N ₂ = 91	37,65	35,84	12,42	12,43

+) Hier sind die Varianzen signifikant voneinander verschieden (F-Test).

Streuung wird hier und im folgenden synonym verwendet mit Standardabweichung (vgl. KREYSZIG 1968, S. 39).

Bei den RW handelt es sich immer um Testrohwerte.

Die USW und GSW stellen Standardwerte mit einem Normmittelwert von 50 dar.

Bei den Stichproben N₁ handelt es sich immer um die jeweils noch vorhandenen Probanden zu einem Untersuchungszeitpunkt, bei N₂ immer um die Ausfallquote zu den in Spalte 1 der Tabelle 3 jeweils aufgeführten Untersuchungssterminen.

Die bisherigen Darlegungen bezogen sich immer auf die Ausgangsstichprobe bzw. deren Schrumpfung. Für die Beurteilung der Relevanz der Untersuchungsergebnisse ist jedoch die Frage der Repräsentativität der Stichprobe nicht ohne Belang (wenngleich für Theorie und Praxis relevante Forschungsergebnisse auch an nicht repräsentativen Stichproben gewonnen werden können - und vice versa), (SADER 1961, HOLZKAMP 1972).

Es besteht daher Veranlassung, die Merkmalsverteilung zu den einzelnen Untersuchungsterminen genauer zu betrachten: Von der Zurückstellung zum Zeitpunkt der Schulpflicht sind die jüngeren Schüler hoch bedeutsam häufiger betroffen als die älteren. Dies gilt für beide Geschlechter, jedoch erheblich mehr für Jungen als für Mädchen. Außerdem ist festzustellen, daß der Anteil der jüngeren Schüler den der älteren überwiegt (nicht signifikant). Ebenso ist die Anzahl der Jungen in unserer Stichprobe größer als die der Mädchen. Wenn man aber bei einer Signifikanzprüfung die Tatsache berücksichtigt, daß das sekundäre Geschlechtsverhältnis in der Bevölkerung 100 Mädchen : 106 Jungen beträgt (FUHRMANN 1965), so ist die Überzahl der Jungen nicht statistisch bedeutsam.

Obwohl die Stichprobe in Tabelle 1 lediglich hinsichtlich dreier Merkmale klassifiziert wurde, sollte doch deutlich geworden sein, daß bei Stichprobenbeschreibungen, die sich in der Literatur häufig nur auf die getrennte - jeweils "eindimensionale" - Angabe der Merk-

malsverteilung beschränkt, für eine angemessene Beurteilung wichtige Zusammenhänge unaufgeklärt bleiben, was zu falschen Schlußfolgerungen führen kann. Daher ist gerade eine Detaildarstellung unumgänglich, wenn

- a) der Interpretationswert von Untersuchungsergebnissen nicht unnötig fragwürdig sein soll,
- b) Literaturvergleiche sinnvoll möglich sein sollen,
- c) bei weiteren Forschungen verwandter Problematik die Möglichkeit gegeben sein soll, auf der Basis vorhandener Ergebnisse präzisere Hypothesen zu formulieren.

Es braucht nicht besonders betont zu werden, daß eine präzisere Stichprobenbeschreibung nicht ausreicht, um die vorgenannten Punkte zu erfüllen, daß z. B. auch die Kenntnis der Analysetechniken, der Datenaufbereitung, etc. dazu notwendig sind.

Sicher wäre es wünschenswert gewesen, die Stichprobe zu den einzelnen Untersuchungszeitpunkten noch nach mehr als nur 3 Merkmalen zu klassifizieren, etwa:

Kindergartenbesuch,

Geschwisterzahl,

Stellung in der Geschwisterreihe,

Konfessionszugehörigkeit,

Schulbildung der Eltern,

Beruf des Vaters,

Berufstätigkeit der Mutter,

Wohnverhältnisse,

Vollständigkeit der Familien,
Einzugsbereich und Ausstattung der Schulen (Image der
Schulen),
Klassenstärken, vor allem in den Grundschulen,
Geschlecht des Lehrers,
Häufigkeit des Lehrerwechsels während der Grundschulzeit,
Art und Dauer der Bewältigung des Schulweges,
um nur einige Klassifikationsmerkmale zu nennen, die alle
"irgendwie" relevant sind für den schulischen Werdegang
und die Entwicklung von Schülern.

Da wir - wie schon gesagt - manche der aufgeführten Merkmale nur unvollständig oder gar nicht erfaßt haben, ist eine derartige Aufschlüsselung hier nicht möglich. Andererseits ist unsere Stichprobe viel zu klein, um derlei sinnvoll durchführen zu können. Dies kann jedoch nicht bedeuten, daß künftig auf derartige Erhebungen und Analysen verzichtet werden dürfe.

Der Schwerpunkt dieser Arbeit liegt in der Analyse der Veränderungen in unserer Stichprobe in der Zeit nach 1964. Daher wird im folgenden eine detailliertere Beschreibung der Stichprobe zum Zeitpunkt der Untersuchung 1968 vorgenommen werden, wobei jeweils alle Klassifikationsmerkmale miteinander permutiert werden.

3.2

Die Längsschnittstichprobe zum Zeitpunkt der Untersuchung 1968
=====

Tabelle 4 enthält eine Zusammenstellung der Stichprobe zum Zeitpunkt der Untersuchung 1968, aufgeschlüsselt nach folgenden Merkmalen:

- 1) dem Einschulungstermin,
- 2) dem Alter in Halbjahren,
- 3) dem Geschlecht,
- 4) dem Beruf des Vaters,
- 5) der besuchten Schulart,
- 6) dem Zeitpunkt des Übergangs auf Real- oder Oberschule,
- 7) den erfolgten Versetzungen resp. Nichtversetzungen.

Wie schon aus Tabelle 1 ersichtlich ist, haben 1968 noch $N = 224$ Pbn an der Untersuchung teilgenommen. Daß unsere Tabelle 4 nur insgesamt 222 Pbn enthält, rührt daher, daß von 2 älteren, altersgemäß eingeschulten, immer versetzten Hauptschuljungen (Vpn: Nr. 218 und 264) der Beruf der Väter nicht zu ermitteln war. Die Angaben im Elternfragebogen lassen auf gestörte Eheverhältnisse schließen. Abgesehen von zwei weiteren Vpn (Vpn 14 und 17), deren Väterberuf dem Elternfragebogen (EMF) entnommen werden mußte, wurde der Beruf des Vaters der Schüler nur nach Angaben im Interessen-Milieu-Fragebogen (IMF) ermittelt.

Bevor wir auf die Verteilung der einzelnen Merkmale eingehen werden, seien noch kurz einige Erläuterungen zur Tabelle 4 gegeben, sofern nicht schon zur Tabelle 1 geschehen:

Tabelle 4: Stichprobe zum Zeitpunkt der Untersuchung 1968

Schulart		SS	HS		RS				OS				
Übergang nach der					4. Klasse		5. Klasse		4. Klasse		5. Klasse		
Versetzung einschl. 30.67			IMMER	NICHT IMMER	IMMER	NICHT IMMER	IMMER	NICHT IMMER	IMMER	NICHT IMMER	IMMER	NICHT IMMER	
Zurückgestellte (Ostern 1961)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4	1										
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3		1									
		4	1	1									
		5	2										
Jüngeres Halbjahr	Mädchen	1											
	2												
	3	2	2	1									
	4	2	7	1		2			1	1			
	5	1	1										
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3		2									
		4		6									
		5	1	8	2	4	3	1	1	1	1		
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3		1		1							
		4		4	3	1							
		5		2		3	1						
Jüngeres Halbjahr	Mädchen	1											
	2												
	3												
	4		4		2								
	5		1	3	2						1	1	
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
	Jüngeres Halbjahr	Mädchen	1										
		2											
		3											
		4											
		5											
altersgemäß Eingeschulte (Ostern 1960)	Älteres Halbjahr	Mädchen	1										
		2											
		3											
		4											</

Erklärungen zu Tabelle 4:

a) Kodierung des Berufs des Vaters (s. Anhang)

- 1 := Berufe, die Hochschulbildung voraussetzen, $N_1=9$
2 := Berufe mit Führungsposition auf mittlerer Ebene, einschl. mittlerer Selbständiger, $N_2=42$
3 := "Stehkragenberufe" mittlerer und unterer Ebene, $N_3=64$
4 := manuelle Berufe (in der Regel Lehrberufe), $N_4=83$
5 := ungelernte Berufe, $N_5=24$

b) besuchte Schulart

- SS := Sonderschüler (Lernbehinderte), $N_{SS}=12$
HS := Hauptschüler, $N_{HS}=100$
RS := Realschüler aller Schulformen, $N_{RS}=50$
OS := Oberschüler aller Schulformen, $N_{OS}=62$

c) Übergang zu einer weiterführenden Schule

- nach Klasse 4 := Schüler, die nach der 4. Klasse Volksschule (Grundschule) zur weiterführenden Schule übergegangen sind,
28 Realschüler, 54 Oberschüler,
nach Klasse 5 := Schüler, die nach der 5. Klasse Volksschule (4 Klassen Grund- und 1 Klasse Hauptschule) zu einer weiterführenden Schule übergegangen sind,
22 Realschüler, 8 Oberschüler.

d) Versetzung bis einschl. Sommer 1967 (max. in Klasse 9)

- IMMER := Schüler, die seit ihrer Einschulung immer in die nächste Klasse versetzt worden sind,
81 Hauptschüler, 36 Realschüler, 44 Oberschüler.

Zu den immer versetzten Schülern zählen auch die wenigen, die von einer "höheren" Schulart in eine "niedere" gewechselt sind, ohne dadurch ein Jahr zu verlieren.

NICHT IMMER := Schüler, die mindestens einmal nicht
versetzt worden sind,
19 Haupt-, 14 Real- und 18 Oberschüler.

Bei den Sonderschülern wurde hier nicht mehr unterschieden.)

Die von uns zur Klassifikation herangezogenen Merkmale
sollten in folgender Hinsicht relevant sein:

1. sollte mittels der gewählten Merkmale Aussagen über
den schulischen Werdegang bzw. das Schulschicksal der
Schüler möglich sein,

2. sollte eine Ineinanderschachtelung rein schulischer und
soziologischer Merkmale Aufschluß über spezifische
Gruppierungen geben,

3. sollte zu allen Klassifikationsmerkmalen eine objek-
tive Zuordnung möglich sein, um Vergleiche möglichst zu
erleichtern. Beim Kriterium "Beruf des Vaters" hängt
die Objektivität jedoch von der Güte unserer Zuordnungs-
skalen ab.

Nun zur Tabelle selbst: Mit dieser von Lilly KEMMLER be-
gonnenen Untersuchung wurde u. W. erstmals in der Bundes-
republik Deutschland ein Einschulungsjahrgang unter-
sucht, der auch die vor oder kurz nach der Einschulung
zurückgestellten Schüler mit einschließt. Wie bereits er-
wähnt, ist es leider nur gelungen, von den Zurückgestell-
ten ca. 2/3 zu erfassen. Ein Grund dafür mag in der Tat-
sache seine Erklärung finden, daß die nicht unerheblichen

zeitlichen und vor allem finanziellen Aufwendungen für die ersten Untersuchungen von Studenten der Psychologie erbracht wurden. Daher wäre es möglich (und verständlich), daß die Beteiligten gerade die bei Nacherhebungen sehr mühevoll Ermittlung der Pbn dann eingestellt haben, als sie glaubten, bei einer bestimmten Quote erreichter Pbn resp. Daten die aufgeworfenen Fragen beantworten zu können. Genau ist dies nachträglich nicht mehr feststellbar.

3.3

Die Klassifikationsmerkmale im einzelnen

Dabei gehen wir so vor, daß wir die einzelnen Merkmale in der angegebenen Reihenfolge einzeln miteinander vergleichen, in der Hoffnung, auf diese Weise eine möglichst gute Übersicht zu gewährleisten.

a) der Einschulungstermin: 1960 oder 1961

Von den insgesamt 224 Schülern wurden 174 (78%) altersgemäß eingeschult und 50 (22%) um ein Jahr zurückgestellt. Die Zurückstellung trifft die Jüngeren erwartungsgemäß hoch signifikant häufiger als die Älteren.

Alter	Einschulung		
	1960	1961	
Jüngere	82	44	126
Ältere	92	6	98
Summe	174	50	224

Die Jungen sind zwar von der Zurückstellung häufiger betroffen als die Mädchen, aber in nicht bedeutsamer Weise.

Alter	Einschulung		
	1960	1961	
Jungen	93	30	123
Mädchen	81	20	101
Summe	174	50	224

Auch hier zeigt sich ein Befund der Bildungsdiskussion nämlich, daß in/um eine(r) Stadt wie Münster mit einem relativ hohen Bildungsangebot schichtspezifische (hier: Beruf des Vaters)⁺⁾ Unterschiede schon bei der Einschulung zu finden sind.

Wie oben bereits angedeutet, wird hier der Beruf des Vaters der Schüler auch identisch mit sozialer Schichtzugehörigkeit verwendet, wohl wissend, hiermit nur einen - wenn auch besonders wichtigen - Schichtindikator zu berücksichtigen.

Für unsere Fragestellung mußten wir uns - wegen der relativ kleinen Stichprobe - auf Zusammenfassung nach der 1. Ziffer beschränken, da andernfalls die Übersichtlichkeit verlorengegangen wäre.

⁺⁾ Einzelheiten über die Kriterien der Berufskodierung finden sich weiter unten sowie bei KEMMLER (1976). Hier sei - vereinfachend - erwähnt, daß wir eine zweiziffrige Kodierung (z. B. 32), 1. Ziffer von 1 - 5, 2. Ziffer von 1 - 6, gewählt haben (Anhang). Nachdem zunächst jeder Beruf nach der 1. Ziffer einer der 5 Gruppen zugeordnet worden war, wurde dort, wo es möglich war, innerhalb dieser Gruppen jeder Proband nach den 6 Kriterien der 2. Ziffer klassifiziert. Dies traf natürlich in erster Linie im Mittelbereich zu; dort bestehen - bekanntermaßen - bei allen in der Literatur verwendeten Berufskodierungen auch die größten Unsicherheiten.

Beruf des Vaters	Einschulung		
	1960	1961	
I	9	0	9
II	31	11	42
III	58	6	64
IV	58	25	83
V	16	8	24
Summe	172	59	222

Kein Schüler, dessen Vater unserer höchsten Berufsgruppe (I = Schicht) angehört, zählt zu den Zurückgestellten, und dies obgleich 8 Probanden dieser Gruppe auch zur Gruppe der Jüngeren gehören. Deutlich ist auch die geringe Quote von 6 Schülern der Gruppe III, die zurückgestellt wurden. Aus den Gruppen II, IV und V wurden dagegen jeweils ca. 30% nicht altersgemäß eingeschult. Faßt man die Gruppen I + II zusammen, so liefert die Gleichverteilungsprüfung eine signifikante Abweichung von den Erwartungswerten, zu "Lasten" der beiden unteren Gruppen.

Als nächstes soll der Einschulungstermin mit der besuchten Schulart verglichen werden.

Schulart	Einschulung		
	1960	1961	
SS	1	11	12
HS	71	29	100
RS	47	3	50
OS	55	7	62
Summe	174	50	224

Augenfällig sind folgende Fakten:

- 1) Die Gefahr, zur Sonderschule zu müssen, ist bei den zurückgestellten Schülern besonders groß (11 der 12 SS gehören zu den Zurückgestellten).
- 2) Die Chance, noch zur RS oder OS überwechseln zu können, ist bei den Zurückgestellten erheblich reduziert.

Die Chi-Quadratprobe dieser Verteilung ist hoch signifikant, d. h. von den Zurückgestellten besuchen mehr Schüler später die SS oder HS als bei Gleichverteilung erwartet werden müßten.

Innerhalb der weiterführenden Schularten RS und OS soll nun untersucht werden, ob es in dieser Substichprobe Unterschiede bezüglich der Grund- oder Hauptschulklasse gibt, nach deren Absolvierung die Schüler die weiterführenden Schularten besuchen.

RS + OS	Einschulung		
	1960	1961	
nach Klasse 4	73	9	82
nach Klasse 5	29	1	30
Summe	102	10	112

Es ergibt sich keine bedeutsame Gleichverteilungsabweichung. Deutlich ist aber, daß die "Weichen" für 73% (82 von 112) der Schüler bereits nach der 4. Klasse gestellt werden.

Schließlich muß das Verhältnis von Immer-Versetzten und Sitzenbleibern noch bezüglich des Einschulungstermins überprüft werden.

Versetzung	Einschulung		
	1960	1961	
IMMER	133	28	161
NICHT IMMER	41	22	63
Summe	174	50	224

Wenn die SS bei dieser Verteilungsprüfung als Sitzenbleiber aufgefaßt werden, so stellen die Zurückgestellten eine sehr signifikant größere Sitzenbleiberquote. Bei Nicht-Berücksichtigung der Sonderschüler ist das Verhältnis in beiden Gruppen etwa ausgeglichen. Insgesamt sind bis zum letzten Untersuchungstermin, Frühjahr 1968, 28% (63 von 224) mindestens einmal "sitzengeblieben".

Zusammenfassend läßt sich sagen, daß die Zurückgestellten im gegenwärtigen Schulsystem Benachteiligungen hinnehmen müssen, die - soweit unsere Angaben reichen, s. u. - nur sehr selten wieder ausgeglichen werden können. Tabelle 4 zeigt augenfällig, daß kein einziger der älteren Zurückgestellten später eine RS oder OS besucht. Die Tatsache, daß auch aus der Gruppe der älteren, altersgemäß eingeschulten Schüler niemand mehr nach der 5. Klasse zur OS überwechselt, läßt vermuten, daß dies keine Zufälligkeiten sind, sondern u. a. bei der Frage des Übergangs auch das zu erwartende Alter bei Schulab-

schluß des betreffenden Schülers in die Überlegungen einbezogen wird.

Früher wie heute gab und gibt es Befürworter der Zurückstellung, wobei das verlorene Jahr als Phase der "Nachreifung" verstanden wurde oder wird, die es dem Schüler anschließend ermöglichen solle, seine Schulausbildung ohne die sonst zu erwartenden Schwierigkeiten zu durchlaufen. Demnach sollten auch unter den Real- und Oberschülern, die 1961 eingeschult wurden, zumindest nicht mehr Versager sein als in der vergleichbaren Gruppe der altersgemäß Eingeschulten.

RS + OS	Einschulung		
	1960	1961	
IMMER VERSETZT	76	4	80
NICHT IMMER VERSETZT	26	6	32
Summe	102	10	112

Tatsächlich zeigt jedoch auch hier die Gleichverteilungsprüfung eine bedeutsame Abweichung zu Lasten der Zurückgestellten. Welchen Einfluß dabei das Faktum hat, daß die Zurückgestellten ein Schulbesuchsjahr und damit einen Versetzungstermin weniger aufweisen, als die altersgemäß eingeschulten Mitschüler, muß einstweilen unberücksichtigt bleiben. Sicher scheint jedoch, daß die Versagerquote unter den Zurückgestellten bei gleicher Schulbesuchsdauer nicht geringer, sondern eher größer wäre. Es bleibt aber festzuhalten, daß von den Zurück-

gestellten, die auf eine weiterführende Schule übergegangen sind, 60% (6 von 10) bis 1968 schon mindestens einmal versagt haben. Unter den Sonder- und Hauptschülern der Zurückgestellten liegt die Versagerquote bei 40% (16 von 40), wobei alle Sonderschüler als Versager aufgefaßt wurden. Insgesamt haben 44% aller Zurückgestellten bis 1968 mindestens einmal die Versetzung nicht erlangt.

Günstiger sind die Verhältnisse bzgl. der Versetzung der altersgemäß Eingeschulten. Bis 1968 haben 20% der HS, 23% der RS und 27% der OS mindestens einmal die Versetzung in die nächste Klasse nicht erreicht. Insgesamt beträgt die Versagerquote der altersgemäß Eingeschulten RS + OS bis 1968 25% (26 von 102). Unter den Sonder- und Hauptschülern liegt diese Quote bei 21% (15 von 72). Die Versagerquote aller altersgemäß Eingeschulten beträgt 24% (von 174).

Bei allen Längsschnittbetrachtungen tritt die wichtige Frage nach den Bezugsgrößen der Stichproben auf. So stellen die hier genannten Versagerquoten Mindestgrößen dar, weil hier - um Verwirrungen zu vermeiden - die Zahl der Schulwechsler nicht berücksichtigt wurde, diejenigen also, die die zunächst "gewählte" weiterführende Schule inzwischen wieder verlassen haben, also von OS zur HS oder HS bzw. von RS zur HS (s. Tabelle 4).

Faßt man Sitzenbleiben, Zurückstellung und verspäteten Übergang zu einer weiterführenden Schule als Behinderung (Überalterung) im schulischen Fortkommen zusammen, so ergibt sich bis einschl. Sommer 1967 - also bis nach der Versetzung in die 9. Klasse - eine Anzahl von (nur) 110 Schülern (49% von 224), die im gegenwärtigen Schulsystem keine zeitlichen Einbußen erlitten haben. Anders formuliert heißt dies, mehr als die Hälfte (51%) unserer 1968 noch erreichten Schüler aller Schularten waren dem gegenwärtigen "System Schule" nicht voll gewachsen. Oder sollte man besser sagen, die Schule war ihnen nicht gewachsen?

Die Zahl der um 2 Jahre Überalterten Schüler der Gesamtstichprobe beträgt zum Zeitpunkt der Untersuchung 1968 schon 12,2% (von 224) - mithin jeder 8. Schüler.

Die Ergebnisse zum Einschulungstermin der Schüler sind derart deutlich, daß eine erste Wertung erlaubt sein mag. Unsere Daten veranschaulichen nachdrücklich, daß die übliche Praxis der Zurückstellung^{+) - ohne individuelle Förderung bis zur Einschulung - nicht zu rechtfertigen ist (vgl. KEMMLER 1967, S. 150; 1976, S. 91-99; LIEDEL 1972). Die Kritik scheint uns auch berechtigt, obgleich uns eine Kontrollgruppe fehlt, eine Stichprobe also, aus der bei Beginn der Schulpflicht niemand zurückgestellt wurde. (Befürworter der heutigen Praxis könnten nämlich einwenden, die Versagerquote sei vermut-}

^{+) Es soll nicht übersehen werden, daß einige Bundesländer z. B. Nordrhein-Westfalen notwendige Konsequenzen gezogen haben.}

lich in der Kontrollstichprobe noch höher.) Ob solch ein Einwand die heutige Praxis rechtfertigen könnte, soll unten noch mit Hilfe der psychologischen Variablen untersucht werden.

b) das Alter (in Halbjahren) der Schüler/innen

Erwartungsgemäß gibt es zwischen Alter und Geschlecht in unserer Stichprobe keine bedeutsamen Unterschiede s.o.

	Jüngere	Ältere	
Jungen	70	53	123
Mädchen	56	45	101
Summe	126	98	224

In der Literatur finden sich zahlreiche (oft allerdings widersprechende) Hinweise dafür, daß die Jahreszeit der Geburt für die Entwicklung eines Neugeborenen nicht ohne Bedeutung sei: So über den Zusammenhang mit Intelligenzentwicklung und Schulleistung (vgl. KEMMLER 1967; KIENER 1969; MARTINDALE/BLACK 1970; DOORNBOS 1971 und 1974; THOMPSON 1971), mit Lernbehinderung (BLACK 1973), mit psychotischen Störungen (DALÉN 1975), mit Extraversion und Neurotizismus (KANEKAR & MUKERJEE 1972), mit sozioökonomischem Status (PASAMANICK/DINITZ/KNOBLOCK 1960; ZELNIK 1969; JAMES 1971; CHAUDHURY 1972) oder mit Lebenserwartung (DANNEEL 1974). Auch mögliche Verbin-

dungen zum Biorhythmus und zur Astrologie werden in ihrem Für und Wider der Beeinflussung der Persönlichkeitsentwicklung erörtert (WENDT 1974 und 1978; PELLEGRINI 1975; MAYO/WHITE/EYSENCK 1978; SMITHERS/COOPER 1978).

Daher soll hier versucht werden, den Geburtstermin unserer Schüler "schicht-" und altersspezifisch etwas genauer zu bestimmen. Von besonderem Interesse ist dies u.E. bei den Gruppen, die erwartungswidrig viele Pbn aufweisen, also bei den $8 + 27 = 35$ jüngeren Schülern unserer Schichten I + II und den 36 Älteren unserer Schicht III. Am Stichtag der ersten Untersuchung, dem 30. 04. 1960, waren alle Schüler zwischen 5;10 - 6;9 Jahre alt. Die Schüler des jüngeren Alteshalbjahres hatten zu diesem Zeitpunkt ein mittleres Alter von 6;0 Jahren, die des Älteren Halbjahres ein mittleres Alter von 6;6 Jahren. Das heißt, der Schwangerschaftsbeginn der Mütter der Gruppe der o.a. 35 Jüngeren lag um den 01. 08. 1954 + 3 Monate, resp. um den 01. 02. 1954 + 3 Monate bei den 36 Älteren.

Die Gleichverteilungsprüfung der beiden Altersgruppen mit den Berufen der Väter unserer Schüler/innen liefert eine signifikante Abweichung; den Berufen I + II entstammen überzufällig viele Kinder, die unserer Gruppe

der Jüngeren angehören, während der Berufsgruppe III entsprechend zu wenige Schüler zur Altersgruppe der Jüngeren zählen, resp. jeweils vice versa, bezogen auf die Gruppe der Älteren. Bei den Berufsgruppen IV und V, Facharbeiter und Ungelernte, ist das Altersverhältnis ausgeglichen.

	Beruf des Vaters					
	I	II	III	IV	V	Summe
Jüngere	8	27	28	51	12	126
Ältere	1	15	36	32	12	96
Summe	9	42	64	83	24	222

Werden die Berufe I und II zusammengefaßt und die Gruppe V weggelassen, so ist die Gleichverteilungsabweichung auf dem 1%-Niveau signifikant.

Daß es in verschiedenen Sozialschichten eine Art von Familienplanung gibt, ist nicht neu. Um die Zusammenhänge, die den Komplex Schulschicksal oder Schullaufbahn von Erstkläßlern betreffen, in diesem Punkt besser kennenzulernen, wäre es wünschenswert zu erfahren, ob unser Befund regional oder/und temporal erklärt werden muß, oder ob hier ein Stichprobenartefakt vorliegt.

Ließe sich unser Befund generalisieren, so müßte die Hypothese, wonach die Entwicklung von Neugeborenen seasonspezifisch verläuft, modifiziert werden, d. h. die bisher beobachteten, unterschiedlichen Entwicklungsverläufe bei Neugeborenen könnten u. U. auf die "Schichtzugehörigkeit" zurückgeführt werden. Dies erscheint nach unseren Daten plausibel, da - wie schon geschildert - erwartungswidrig viele Kinder unserer beiden oberen "Schichten" im Frühjahr geboren wurden, dem Zeitpunkt, der für die Entwicklung Neugeborener als besonders günstig bezeichnet wird.

Weitergehende Erkenntnisse hierzu dürften von einiger Bedeutung sein für die Festlegung des Einschulungstermins bei Beginn der Schulpflicht (vgl. TIETZE 1973). So könnte z. B. die Verlegung des Einschulungstermins in der Bundesrepublik Deutschland vom 01. April auf den 01. August im Jahre 1966 Einfluß auf die Altersstruktur der Schüler haben, die nach der Umstellung eine weiterführende Schule besuchen. Ebenso könnte festgestellt werden, ob nach der Neufestsetzung des Einschulungstermins die jüngeren Schüler wieder häufiger von der Zurückstellung betroffen sind als die älteren und welcher Sozialschicht sie angehören. Dazu bedarf es natürlich eines Fixpunktes, von dem aus die Alterseinteilung - etwa in Halbjahre - vorgenommen wird. Dies alles wäre ohne großen Aufwand prüfbar, würde aber den Rahmen dieser Arbeit verlassen.

Als weiteres Kriterium ist die besuchte Schulart mit dem Alter der Schüler zu vergleichen, wobei jedoch nur - erwartungsgemäß - zufällige Abweichungen nachweisbar sind.

	Schulart				
	SS	HS	RS	OS	
Jüngere	7	58	22	39	126
Ältere	5	42	28	23	98
Summe	12	100	50	62	224

Im folgenden soll der Zeitpunkt des Übergangs der weiterführenden Schüler zu einer RS oder OS zum Alter in Beziehung gebracht werden.

Übergang	Jüngere	Ältere	
nach Klasse 4	40	42	82
nach Klasse 5	21	9	30
Summe	61	51	112

Die Gleichverteilungsabweichung ist tendenziell bedeutsam ($p < 10\%$). Dies besagt; unter den Jüngeren gehen mehr Schüler erst nach der 5. Klasse zu einer weiterführenden Schule.

Schließlich sind unsere beiden Altersgruppen noch hinsichtlich der Versetzung zu kontrollieren. Abweichungen von der Gleichverteilung sind hier nur zufällig.

Versetzung	Jüngere	Ältere	
IMMER	91	70	161
NICHT IMMER	35	28	63
Summe	126	98	224

c) Geschlecht der Schüler/innen

Zunächst sollen die Relationen Beruf des Vaters (Schichtzugehörigkeit) und Geschlecht der Schüler auf Zufälligkeit hinsichtlich Abweichungen von der Gleichverteilung überprüft werden.

Beruf des Vaters	Jungen	Mädchen	
I	6	3	9
II	24	18	42
III	35	29	64
IV	46	37	83
V	10	14	24
Summe	121	101	222

Die Verteilungsabweichungen sind - erwartungsgemäß - nur zufällig von Null verschieden. Nicht so beim Vergleich der Geschlechter mit der besuchten Schulart:

Schulart	Jungen	Mädchen	
SS	8	4	12
HS	52	48	100
RS	21	29	50
OS	42	20	62
Summe	123	101	224

Die Gleichverteilungsprüfung ist signifikant, d. h. die ungleiche Verteilung geht bei den Sonder- und Oberschülern zu "Lasten" der Jungen, bei den Realschülern zu "Lasten" der Mädchen. Wenn man die Verteilungsprüfung nur für die RS und OS durchführt, so ist das Ergebnis in der oben beschriebenen Weise sehr signifikant.

Desweiteren soll nun überprüft werden, ob sich die Verteilungen der Pbn hinsichtlich Geschlecht sowie weiterführende Schulart (RS/OS) und der Klasse, nach der die betreffende Schulart gewählt wurde, überzufällig voneinander unterscheiden.

Schulart	Übergang	Jungen	Mädchen	
RS	nach Klasse 4	13	15	28
	nach Klasse 5	8	14	22
OS	nach Klasse 4	36	18	54
	nach Klasse 5	6	2	8
Summe		63	49	112

Die Verteilungen weichen bedeutsam voneinander ab; Jungen gehen erwartungswidrig selten nach der 4. Klasse zur RS, jedoch erwartungswidrig häufig nach der 4. Klasse zur OS. Das Ergebnis besagt: Die Entscheidung für den Besuch einer Oberschule wird für Jungen eher gefällt, als die für den Besuch einer Realschule und vice versa für Mädchen. Mögliche Ursachen hierfür müssen später im Zusammenhang mit der Analyse der Testergebnisse referiert werden. Es sei hier hervorgehoben, daß es anscheinend

"leichter" fällt, jemandes Oberschuleignung zu bestimmen, obgleich die Anforderungen an die Oberschüler (vermutlich) größer sind als dies bei Realschülern der Fall ist, und obgleich die reguläre Schulbesuchszeit 3 Jahre länger währt, sofern der Eignungsfrage bei dieser Entscheidung überhaupt besondere Bedeutung zukommt.

Abschließend zu dem Vergleich der Geschlechter mit den übrigen Klassifikationskriterien sei hinzugefügt, daß Jungen und Mädchen vom Kriterium Versetzen/Sitzenbleiben gleichermaßen betroffen sind.

Versetzung	Jungen	Mädchen	
IMMER	86	75	161
NICHT IMMER	37	26	63
Summe	123	101	224

d) Beruf des Vaters der Schüler/innen

Als erstes sollen der Beruf des Vaters der Schüler (soziale Schicht) und die besuchte Schulart auf Gleichverteilung überprüft werden.

Schulart	Beruf des Vaters					
	I	II	III	IV	V	
SS	0	2	0	4	6	12
HS	0	14	25	46	13	98
RS	1	11	16	18	4	50
OS	8	15	23	15	1	62
Summe	9	42	64	83	24	222

Es fällt auf, daß aus Gruppe I niemand zur Sonder- oder Hauptschule und auch lediglich ein Schüler zur Realschule geht. Außerdem ist sicher nicht unwesentlich, daß aus der relativ großen Gruppe III kein Schüler die Sonderschule besucht. Faßt man aus formalen Gründen die Gruppen I + II sowie die SS + HS zu jeweils einer neuen Gruppe zusammen, so daß eine $4 \times 3 = 12$ Feldertafel entsteht, so liefert die Verteilungsprüfung ein hoch signifikantes Resultat zu "Lasten" der Gruppen IV und V, d. h. in diesen Sozialschichten bestehen ganz erhebliche Bildungsebenachteiligungen.

Als nächstes soll geprüft werden, ob der Beruf des Vaters einen Einfluß darauf hat, wann ein Kind zu einer weiterführenden Schule übergeht.

Übergang zur RS/DS	Beruf des Vaters					
	I	II	III	IV	V	
nach Klasse 4	7	14	30	26	5	82
nach Klasse 5	2	12	9	7	0	30
Summe	9	26	39	33	5	112

Faßt man die Gruppen IV und V zusammen, so wird die Chi²-Prüfung auf dem 10%-Niveau signifikant, was hauptsächlich zu "Lasten" der Gruppe II geht. Es scheint die Tendenz vorzuherrschen, entweder ein Kind möglichst früh zu einer weiterführenden Schule zu schicken oder gar nicht.

Abschließend auch hier wieder der Vergleich des Referenzkriteriums mit der Versetzung der Schüler, hier wie bei den vorhergehenden und nachfolgenden Abschnitten der Stichprobenbeschreibung bezogen auf den Zeitpunkt der Untersuchung (Frühjahr) 1968, also nach der Versetzung der am weitesten Fortgeschrittenen in die 9. Klasse.

Versetzung	Beruf des Vaters					
	I	II	III	IV	V	
IMMER	7	31	45	61	15	159
NICHT IMMER	2	11	19	22	9	63
Summe	9	42	64	83	24	222

Die Abweichungen von der Gleichverteilung sind nur zufällig von Null verschieden.

Dies ändert sich jedoch, wenn die Gruppe der OS allein betrachtet wird.

Versetzung	Beruf des Vaters			
	I + II	III	IV + V	
IMMER	18	12	14	44
NICHT IMMER	5	11	2	18
Summe	23	23	16	62

Für diese Gruppe ergibt sich eine signifikante Abweichung von der Gleichverteilung, d.h. die Schüler der "Schichten" I + II und IV + V bleiben weniger häufig sitzen als die-

jenigen der "Schicht" III. Es wäre wünschenswert zu wissen, ob speziell bei den Schülern aus IV + V die Übergangsbedingungen schon eine spezifische Rolle spielen: Mit anderen Worten: Ist dies Ergebnis so zu interpretieren, daß beim Übergang aus "Schicht" IV + V eher der zu erwartende Erfolg (vielleicht aus der Sicht des Lehrers) vorrangig ist, während bei den Schülern aus III der Elternwunsch beim Übergang von größerer Bedeutung ist? Unklar ist auch die Ursache für die geringere Sitzenbleiberquote der "Schichten" I und II (vgl. KEMMLER 1976).

e) Besuchte Schularart der Schüler/innen

Bevor wir in der Darstellung der Stichprobe unter bestimmten Klassifikationsmerkmalen fortfahren, seien einige übergeordnete Ausführungen zur Zusammensetzung der Stichprobe angeführt.

In unserer Stichprobe sind die 4 wesentlichen allgemeinbildenden Schularten vertreten. Die Schüler befanden sich zum Zeitpunkt der Untersuchung 1968 (Stichtag: 15. 02. 68) in/auf insgesamt 34 Schulen:

- 2 Sonderschulen (5,4% von 224 Schülern,
- 17 Hauptschulen (44,6%),
- 7 Realschulen (22,3%),
- 8 Oberschulen (27,7%).

Bei Analyse der Testdaten werden wir mitzudiskutieren haben, wieweit Varianzunterschiede innerhalb und zwischen den Schularten bedingt sind oder sein können durch den Besuch bestimmter Schulen innerhalb der 4 Schularten (spezifische Schul- und/oder Klasseneffekte).

Bei den vorher behandelten Kriterien war auch immer schon von der besuchten Schulart die Rede. Hier bleibt zu klären, ob der Übergang zu einer weiterführenden Schule abhängig ist von der Klasse, nach der dieser Übergang vollzogen wird.

Übergang	Schulart		
	RS	OS	
nach Klasse 4	28	54	82
nach Klasse 5	22	8	30
Summe	50	62	112

Die Verteilungsprüfung zeigt, daß die RS gegenüber den OS hoch signifikant häufiger erst nach der 5. Klasse zur Realschule überwechseln, bzw. daß die OS hoch bedeutsam häufiger schon nach der 4. Klasse die Oberschule besuchen. Es bestätigen sich hier noch einmal die Befunde, die wir auf Seite 38 und 39 in anderem Zusammenhang gewonnen haben.

Bleibt noch zu prüfen, ob die Klassifizierung der Stichprobe nach besuchter Schulart und Versetzung/Nicht-Versetzung zu erwartungswidrigen Verteilungsabweichungen führt.

Versetzung	Schulart				
	SS	HS	RS	OS	
IMMER	0	81	36	44	161
NICHT IMMER	12	19	14	18	63
Summe	12	100	50	62	224

Bedeutsame Verteilungsabweichungen sind nicht nachweisbar.

f) Übergang zur Real- oder Oberschule

Nachdem wir die in Tabelle 4 für die Beschreibung und Prüfung der Stichprobe verwendeten Klassifikationsmerkmale permutiert haben, bleibt letztlich nur noch zu prüfen, ob es spezifische Unterschiede gibt zwischen dem Zeitpunkt des Übergangs zur RS oder OS und der Versetzung/Nicht-Versetzung.

Versetzung	Real- oder Oberschule		
	nach Klasse 4	nach Klasse 5	
IMMER	57	23	80
NICHT IMMER	25	7	32
Summe	82	30	112

Die Chi²-Probe weicht nur zufällig von Null ab.

.0 DIE VERWENDETEN PSYCHOLOGISCHEN VERFAHREN

.1 Ausgangsuntersuchung Ostern 1960 bzw. Ostern 1961

Die Auswahl der verwendeten Verfahren richtete sich logischerweise nach den Intentionen zu den einzelnen Untersuchungszeitpunkten (s. Überblick). Da zunächst die Frage der Schulreife zum Zeitpunkt der Einschulung in die Grundschule im Vordergrund des Interesses stand, wurden bei den Untersuchungen 1960/61 die nachfolgend aufgeführten Verfahren durchgeführt. Unsere Darstellung wurde weitgehend von KEMMLER (1967, S. 31f) übernommen.

1. Grundleistungstest von KERN (1957, 1963)

Es ist ein sprachfreies Verfahren und besteht aus sechs Untertests, die verschiedene Aspekte der "Gliederungsfähigkeit" messen sollen. In der Schulpraxis wird - wie in dieser Untersuchung - meist eine gekürzte Form mit drei Aufgaben verwendet:

- a) Die zeichnerische Darstellung eines Kindes.
- b) Das Nachmalen eines einfachen Satzes (z. B. "Udo ist da").
- c) Die zeichnerische Wiedergabe einer vorgegebenen Mengengestalt, z. B.

. . .
. . . .
. . .

2. Auslesetest für Schulneulinge, Form B, von WALTER (o.J.) ("Münchener Schulreifetest")

Dieser sprachfreie Gruppentest enthält sechs Untertests:

- (1) Formauffassung, (2) Formwiedergabe, (3) Formaufbau,
 (4) Rechenbereitschaft, (5) Sprach- und Situationserfassung,
 (6) Freies Zeichnen. Die Untertests 1, 2 und 6 fordern

Überblick über die einzelnen Untersuchungen

Untersuchungs- zeitpunkt	Probanden	Verfahren
1960/61	N = 321	.Grundleistungstest .Auslesetest f. Schulneulinge .Göppinger Schulreife-test .Fragebogen zur seelisch- sozialen Schulreife
	ausgelesene Gruppe	.BINET-BOBERTAG-Intelli- genztest
1964	N = 251	PMA 8 - 10: Bilderwortschatztest Wortschatztest Raumerfassungstest Wortgruppentest Figurengruppentest Wahrnehmungsschnelligk. Zahlentest Wissenstest Wörtertrennentest Wörtersuchentest Fehlersuchentest Lehrerfragebogen Kinderfragebogen
1968	N = 224	PMA 13/14: Wortschatztest Figurenvergleichstest Buchstabenreihentest Wortgruppentest Wahrnehmungsschnelligk. Zahlentest Wortanalogientest Textaufgabentest Figurengruppentest Konzentrationsleistungstest Interessen-Milieu-Frageb. IAR-Fragebogen ENRA-Fragebogen Lehrerfragebogen Elternfragebogen
1972	N = 46 (nur OS)	IST'70 (AMTHAUER) Schülerfragebogen

die Fähigkeit zum Erkennen und Wiedergeben anschaulicher Formen nach einer Vorlage oder aus dem Gedächtnis; in Untertest 3 sind die Aufbauelemente eines zu schaffenden Bildganzen zu entdecken, Untertest 4 prüft die Rechenbereitschaft an dargestellten Mengen; Untertest 5: Szenen aus einmal vorgesprochenen kleinen Erzählungen und Kinderreimen müssen unter einer größeren Anzahl von Bildern wiedergefunden werden.

3. Göppinger Schulreifetest von KLEINER (1959) -
(wurde nur bei der Untersuchung 1960 verwendet)

Dieser ebenfalls sprachfreie Gruppentest enthält zehn Untertests (s. auch KEMMLER & KIEKHEBEN, 1962).

Für jedes der drei Schulreifeverfahren werden die richtigen Lösungen in jedem Untertest zu einem Gesamtwert addiert. Die Testautoren geben kritische Punktwerte an, bei denen Schulunreife oder Schulreife vorliegt.

4. Fragebogen zur seelisch-sozialen Schulreife (HECKHAUSEN & KEMMLER, 1957)

Er enthält fünfzehn Fragen zu den folgenden Verhaltenskreisen: (1) Ablösung des Kindes vom Elternhaus; (2) Kontaktverhalten zu Gleichaltrigen; (3) Selbständigkeit; (4) Einordnung in die Unterrichtssituation; (5) Anfänge erkennbarer Leistungsbemühungen (s. KEMMLER, 1967, S. 198).

5. Mit allen nach den Schulreifetests "nicht-schulreifen" sowie "fraglich-schulreifen" und allen zurückgestellten Schülern wurde im Einzelversuch der Intelligenztest von

SIMON-BOBERTAG (in der Bearbeitung von I. NORDEN, 1953) durchgeführt, um so möglichst spezifische Hinweise zwischen Schulreife und Intelligenz zu erhalten.

.2

Untersuchung Frühjahr 1964

Bei der Darstellung der verwendeten Verfahren und der Untersuchungsabsicht folgen wir weitgehend KEMMLER (1967, S. 18ff, S. 30):

.2.1

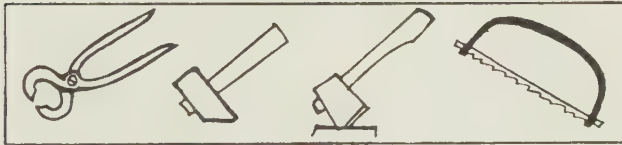
Intelligenztest (Testbatterie)

KEMMLER wollte einen Intelligenztest verwenden, der für jeden Probanden die Bestimmung eines oder mehrerer Kennwerte der geistigen Leistungsfähigkeit - Standardwert, Intelligenzquotient - ermöglicht. Mit einem solchen, aus mehreren Untertests berechneten Kennwert ist eine brauchbare Kennzeichnung der Intelligenzhöhe der einzelnen Schüler erreichbar. Damit besteht die Möglichkeit, alle anderen Befunde mit diesem Wert in Beziehung zu setzen und Vergleiche mit den Untersuchungsbefunden anderer Autoren anzustellen. Diese Testbatterie sollte möglichst aus solchen Subtests bestehen, die es erlauben, ein Intelligenzprofil zu zeichnen.

Man entschied sich für die "Primary Mental Abilities" von THURSTONE (1938, 1954). Dieser Test ist nach einem faktorenanalytischen Modell konstruiert worden. Er prüft nach Aussage seines Autors mit seinen sieben Untertests insgesamt fünf unabhängige Faktoren (s. auch KEMMLER und LANGHEINRICH 1967; ROTERS 1966).

(1) Bilderwortschatztest

Dieser Test besteht aus 37 Reihen (Aufgaben) zu je vier Bildern. Der Untersuchungsleiter nennt ein Wort und der Proband muß das Bild durchstreichen, das dem Wort entspricht, z. B. 5. Reihe: Sucht die Zange!



Anzahl der Items: 37

(2) Wortschatztest

Bei diesem Test muß aus je vier Wahlwörtern dasjenige angestrichen werden, das die gleiche oder eine sehr ähnliche Bedeutung hat wie das vorangestellte Wort, z. B.

entfernen wegnehmen behalten besitzen stehlen

Hinterhalt Krieg Falle Zaun Feind

Anzahl der Items: 40

Beide Tests messen nach THURSTONE den Faktor V - verbal Meaning -; die Fähigkeit, Begriffe, die in Worten ausgedrückt sind, zu verstehen.

(3) Raumerfassungstest

Hier gibt es in jeder Reihe vier Auswahlfiguren, aus denen eine herausgesucht werden muß, die die vorangestellte Figur zu einem Quadrat ergänzt, z. B.



Anzahl der Items: 27

Dieser Test mißt nach Meinung des Testautors den Faktor S - space -; die Fähigkeit, sich zwei- oder dreidimensionale Gebilde vorzustellen und damit umzugehen.

(4) Wortgruppentest

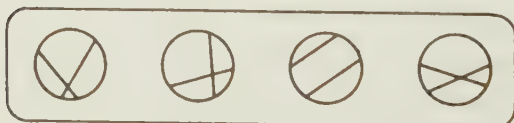
Unter den vier Wörtern einer Reihe muß dasjenige herausgefunden werden, das nicht in die Reihe paßt, z. B.

klein	freundlich	gut	glücklich
Schaufel	Schere	Säge	Messer

Anzahl der Items: 40

(5) Figurengruppentest

In jeder Reihe stimmen von vier Konfigurationen drei in ihrer Lage, Richtung, Zuordnung Elemente usw. überein. Die nicht in die Reihe passende Konfiguration muß durchgestrichen werden, z. B.



Anzahl der Items: 27

Untertest 4 und 5 messen den Faktor R - reasoning -; die Fähigkeit, Probleme logisch zu lösen, Einsicht zu nehmen, zu planen.

(6) Wahrnehmungsgeschwindigkeitstest

In jeder Reihe dieses Untertests befinden sich vier gegliederte Gebilde, von denen jeweils zwei genau gleich sind und zwei von den anderen in einigen Details geringfügig abweichen. Hier sind die gleichen Gebilde in jeder Reihe möglichst schnell herauszufinden und zu markieren, z. B.



Anzahl der Items: 50

Dieser Test soll den Faktor P - perceptual speed - messen; die Fähigkeit, Gleichheit und Unterschiede zwischen Objekten oder Symbolen schnell und genau zu erkennen.

(7) Zahlentest

Hier wird verlangt, Additionsaufgaben mit zwei oder drei Summanden aus höchstens zweistelligen Zahlen zu lösen, z. B.

$$\begin{array}{r} 62 \\ 45 \\ \hline \end{array} \quad \begin{array}{r} 2 \\ 3 \\ 2 \\ \hline \end{array}$$

Anzahl der Items: 52

Dieser Subtest soll den Faktor N - number - messen; die

Fähigkeit, mit Zahlen umzugehen und einfache Rechenaufgaben richtig zu lösen.

(8) Wissenstest

Weiter sollte die Wissensbreite der Schüler abgeschätzt werden, da z. B. nach WECHSLER "der Wissensumfang eines Menschen im allgemeinen ein sehr gutes Anzeichen seiner intellektuellen Fähigkeit" ist (1956, S. 90). Die Anregung zu Aufgaben für einen Wissenstest wurde teils amerikanischen Verfahren entnommen, teils dem H a w i k (Bern 1956), einige Aufgaben wurden selbst entworfen (KEMMLER und LANGHEINRICH 1969).

Jede Aufgabenreihe besteht aus einer Frage und vier Alternativantworten, von denen die richtige markiert werden muß, z. B.

Wasser ist schwerer als?	Blei	Öl	Eisen	Stein
Ein Dutzend hat?	12 Stck	6 Stck	50 Stck	12 Stck
Anzahl der Items: 39				

(9) Wörtertrennentest

Hierbei wurde ein Diktat aus dem Schulleistungstest von WITTMANN (1964) verwendet, das für das 3. Schuljahr gedacht ist.

Auf den Testbögen ist der Wortlaut des Diktates in Großbuchstaben so geschrieben, daß der übliche Zwischenraum zwischen den einzelnen Wörtern wegfällt; alle Buchstaben haben also den gleichen Abstand voneinander. Außerdem gehen die Wörter nach dem Ende einer Zeile ohne Rücksicht auf die Trennungsregeln in der nächsten Zeile weiter.

Die ersten beiden Zeilen auf dem Testblatt:

W I R R A S T E N A U F E I N E R W I E S E . D O R
T S T E H T E I N B A U M . U N T E R I H M S I T Z

Die Probanden werden aufgefordert, hinter jedes Wort einen senkrechten Strich zu machen.

Augenscheinlich prüft dieses Verfahren Gliederungs- und Lesefähigkeit.

Anzahl der Items: 53

(10) Wörtersuchentest

Einfache, den Schülern bekannte Wörter aus vier Buchstaben sind in einem längeren Text aus sonst sinnlosen Buchstabenfolgen verstreut, z. B.

B E R G N X A D O R F E I M O N D K L I T H A U S A
C I B D A C H K F F L H E R Z U S R N K M N E U N W
A I G B E T T L N D N I A U G E B Y D O T T P O R N

Die Kinder sollen die sinnvollen Wörter herausuchen und mit dem Stift einkreisen.

Es hat sich gezeigt, daß sich die Testleistungen - das Herauslösen von Wortgestalten aus einer Zufallsfolge von Buchstaben - sehr gut zur Differenzierung zwischen guten und schwachen Leser eignen.

Anzahl der Items: 46

(11) Fehlersuchentest

Bei diesem Test handelt es sich um den Untertest 14 des Leistungsprüfsystem (LPS) von HORN (1962), Form A.

Der Test soll die Wahrnehmungsschnelligkeit der Probanden prüfen. Er wurde - außer dem bereits erwähnten Test - in die Untersuchung aufgenommen, um die Fähigkeit im Faktor Wahrnehmungsschnelligkeit, sensu THURSTONE, besser abschätzen zu können.

Die Schüler haben je Aufgabe eine Konfiguration aus numerischen oder alpha-numerischen Zeichen auf Kongruenz mit einer zweiten zu prüfen. Inkongruenzen sind genau zu markieren, z. B.

102001	102008	37029641	37029241
T73N16	M73N16	NLS658AL	NLS668AL

Anzahl der Items: 60

4.2.2

Fragebögen

(12) Kinderfragebogen

Der Kinderfragebogen sollte Fragen nach der Schulanpassung, dem Verhältnis zu anderen Menschen und zu sich selbst, Fragen zur Erfassung der Angst und des Neurotizismus sowie Fragen nach den Interessenrichtungen der Schüler enthalten. Die Konstruktion eines eigenen Interessenfragebogens wurde fallengelassen, da die Interessenrichtungen neunjähriger Kinder nicht ausgeprägt genug schienen.

Beispiele:

Machst du gerne Schularbeiten?

☐ — ☐ — ☐ — ☐ — ☐
sehr gern mittel ungern sehr
gern ungern

Bist du bei Klassenarbeiten aufgeregt?

☐ — ☐ — ☐ — ☐ — ☐
immer oft manchmal selten nie

Die zutreffende Antwortmöglichkeit muß angekreuzt werden (vgl. KEMMLER 1967, S. 186ff).

Anzahl der Items: 77

Für einige besonders interessierende Persönlichkeitsbereiche sollten aus den entsprechenden Fragen Punktwertgruppen gebildet werden, um eine sicherere Beurteilung bestimmter Persönlichkeitsmerkmale zu ermöglichen. Als Denkhilfe wurde dazu das mathematisch-statistische Verfahren der Faktorenanalyse verwendet.

Die Punktwertgruppen wurden unter Zuhilfenahme der Querschnittuntersuchung von KEMMLER (1967, S. 14ff) gebildet. Insgesamt wurden über die Kinderfragebogenerhebung der Querschnittuntersuchung drei Faktorenanalysen berechnet: Über die Extremstichproben "leistungsstarke Gesamt", leistungsschwache Jungen" und leistungsschwache Mädchen".

Es wurden sieben Punktwertgruppen gebildet, die 33 der 77 Einzelfragen enthalten. Die Fragen 23 und 44 finden sich in jeweils zwei Punktwertgruppen.

Tabelle 5 : Zusammensetzung der Punktwertgruppen und die darin eingegangenen Fragen

Punktwertgruppe	lfd. Nr. der Fragen
Verhältnis zur Schulleistung	6, 7, 14, 15, 23
Konzentration	44, 51, 55, 57
Schulanpassung	12, 17, 20, 23, 30, 44
Beziehung zum Lehrer	13, 32, 35, 60, 65, 69
Kontakt	22, 25, 36, 42, 53
Affektivität	34, 52, 70
Angst	2, 28, 31, 64, 72, 74

Bis auf die Gruppe "Kontakt" finden sich alle Punkt-
wertgruppen in allen 3 Faktorenanalysen wieder. Die
Gruppe "Kontakt", die nur in der Faktorenanalyse
"leistungsstarke Gesamt" eine durchgehende Entsprechung
hat, wurde allein nach dem Inhalt der Fragen gebildet
(vgl. NIEMEIER 1968).

(13) Lehrerfragebogen zum Leistungs- und Sozialverhalten
im 4. Schuljahr

Dieser den Lehrern vorgelegte Fragebogen mit Items zum Leistungs- und Sozialverhalten der Schüler ist analog dem im ersten Schuljahr vorgelegten konstruiert, wobei jedoch die Fragen dem Alter der Schüler angepaßt wurden. (s. KEMMLER, 1967, S. 199), z. G.

Hat es (das Kind) Kontakt zu anderen Kindern?

□ — □ — □ — □ — □
zu sehr vielen manchen einem keinem
vielen (3 - 4)

Erwartet es bei Aufgaben Hilfe von Ihnen?

□ — □ — □ — □ — □
nie selten manchmal häufig immer

Auch hier muß die zutreffende Antwortmöglichkeit angekreuzt werden.

Anzahl der Items: 11

4.3

Untersuchung Frühjahr 1968

Eine der wichtigsten Fragen dieser Untersuchung bezog sich auf die Stabilität derjenigen psychologischen Merkmale, die bereits Gegenstand der vorausgegangenen Untersuchungen gewesen waren. Darüber hinaus sollten auch diesmal einige spezielle Begabungen geprüft werden, die im Zusammenhang mit Schulerfolg stehen.

Von den Schülern wurden deren Interessen, Berufswünsche und Bereiche des häuslichen Milieus erfragt. Außerdem wurden ihnen Fragebögen vorgelegt, die Auskunft über Neurotizismus vs. Extraversion und die Einstellung zur Verantwortlichkeit bei eigenem Erfolg oder Mißerfolg

geben sollten.

Den Lehrern wurde wieder ein Fragebogen zum Leistungs- und Sozialverhalten der nunmehr 13 - 14jährigen Schüler vorgelegt. Dieser Fragebogen enthielt auch Fragen zum wahrscheinlichen Schulabschluß der Schüler innerhalb des allgemeinbildenden Schulwesens.

Schließlich wurde den Eltern der Schüler ein Fragebogen zugeschickt, um von ihnen Auskünfte über die Familienverhältnisse zu erlangen; etwa Schulbildung der Eltern, Beruf(tätigkeit), Wohnverhältnisse, Einkommen, etc.

Einige dieser Fragen sollten eine Kontrollfunktion zu den den Schülern gestellten Fragen zukommen.

Die Verfahren im einzelnen:

4.3.1

Intelligenztest (Testbatterie)

Für den diesmal zu verwendenden Intelligenztest galt das Gleiche wie für die 1964 verwendete Version. Darüber hinaus sollte jedoch eine Aussage über die Stabilität intellektueller Fähigkeiten möglich sein. KEMMLER und ZINGELER entschieden sich daher dazu, eine Testbatterie analog den 1964 benutzten Primary Mental Abilities (PMA 8 - 10) zusammenzustellen, die schließlich in Anlehnung an THURSTONE's PMA 11 - 17 als PMA 13 - 14 aus folgenden sechs Untertests bestand (vgl. REICHEL 1969):

(1) Wortschatztest

Unter einem unterstrichenen Schlüsselwort stehen fünf Wörter. Es soll dasjenige Wort herausgefunden und durchgestrichen werden, das dem Schlüsselwort möglichst synonym ist oder am besten zu ihm paßt, z. B.

lauschen

laufen
tauschen
schwätzen
horchen
hauchen

Barke

Rinde
Gerät
Boot
Schwiele
Dampfer

Anzahl der Items: 50

Die Items wurden dem Frankfurter-Wortschatztest (WST 7 - 8, Form A, für 13 - 15 Jahre) und dem Wilde-Intelligenztest (WIT, Form 1, Untertest: Gleiche Wortbedeutung) entnommen. Der Test soll eine wesentliche Komponente des Faktors V - verbal meaning - sensu THURSTONE messen.

(2) Figurenvergleichstest

Von sechs Auswahlfiguren sollen die zwei oder höchstens drei herausgefunden und durchgestrichen werden, die mit der jeweils vorangestellten durch Drehung in der Ebene zur Deckung gebracht werden können, z. B.



Anzahl der Items: 20

Dieser Test wurde den PMA 11 - 17 von THURSTONE (1958) entnommen und soll den Faktor S - space - messen, d. h. die Fähigkeit, sich räumliche Gebilde vorzustellen und damit umzugehen.

(3) Buchstabenreihentest

Jede Buchstabenreihe (Aufgabe) ist nach einer bestimmten Regel aufgebaut. Sie soll gefunden und die Aufgabe im Sinne dieser Regel um einen Buchstaben fortgesetzt werden, den man aus der für jede Reihe zur Verfügung stehenden Buchstabengruppe auswählt und dort durchstreicht, z. B.

a b c c b a d e f f e d g h i / h i j k l m

a b c b c d e f e f g h i h / f g h i j k

Anzahl der Items: 30

Der Test wurde den PMA 11 - 17 von THURSTONE entnommen.

(4) Wortgruppentest

Jede Aufgabe dieses Tests besteht aus fünf Wörtern, die in einer Zeile angeordnet sind, von denen vier einen gemeinsamen Oberbegriff haben. Das nicht zu diesem Oberbegriff passende Wort soll herausgefunden und angekreuzt werden, z. B.

Schaufel	Hobel	Schere	Säge	Messer
hinten	hervor	hinab	seitwärts	hinein

Anzahl der Items: 40

Der Test wurde aus Aufgaben des Wilde-Intelligenztests und des Intelligenz-Struktur-Tests (IST, von AMTHAUER zusammengestellt. Der Wortgruppen- und der Buchstabenreihentest sollen den THURSTONE'schen Faktor R - reasoning - die Fähigkeit des schlußfolgernden Denkens messen.

(5) Wahrnehmungsschnelligkeitstest

Dieser Test ist identisch mit dem gleichnamigen, 1964 verwendeten Verfahren. Die Bearbeitungszeit wurde jedoch reduziert.

(6) Zahlentest

Als Zahlentest wurde der Konzentrations-Leistungs-Test (KLT) von DÜKER & LIENERT (1959) verwendet, wobei jedoch nur die Ergebnisse der ersten sieben Minuten benutzt wurden.

Aufgabenbeispiele:

$$8 + 9 - 2 \quad \underline{\hspace{2cm}} \qquad 3 + 6 - 8 \quad \underline{\hspace{2cm}}$$

$$5 - 4 + 3 \quad \underline{\hspace{2cm}} \qquad 9 + 1 + 7 \quad \underline{\hspace{2cm}}$$

Es soll die obere Zeile ausgerechnet und das Ergebnis gemerkt werden. Dann wird die untere Zeile ausgerechnet und anschließend das kleinere von größeren Ergebnis subtrahiert. Die Differenz wird in das nebenstehende Kästchen eingetragen.

Dieser Untertest soll den Faktor N - number -, die Fähigkeit mit Zahlen umzugehen und einfache Rechenaufgaben schnell und richtig zu lösen, messen.

4.3.2

Spezielle Fähigkeiten

(7) Konzentrationsfähigkeitstest

Verwendet wurde der bereits erwähnte KLT von DÜKER & LIENERT in der Form für Volksschüler, wie als Zahlentest beschrieben. Um den Test nicht zweimal durchzuführen, was auch testtheoretische Konsequenzen gehabt hätte, wurden die Probanden nach sieben Minuten Bearbeitungszeit aufgefordert, unter die zuletzt gelöste Aufgabe einen Querstrich zu machen und anschließend gleich weiterzurechnen (insgesamt 30 Minuten).

Anzahl der Items: 250

(8) Wortanalogientest

Zwischen jeweils zwei vorgegebenen Wörtern besteht eine bestimmte Beziehung. Eine ähnliche Relation besteht auch zwischen dem dritten Wort und einem von fünf Lösungswörtern. Dieses Wort soll herausgefunden und durchgestrichen werden, z. B.

Stadt : Dorf = See : ?

Teich Ort Fluß Land Wasser

Überlegen : entscheiden = Gerichtsverhandlung : ?

Beratung Vergehen Anklage Urteil Verteidigung

Anzahl der Items: 48

Die Aufgaben entstammen größtenteils dem Frankfurter - Analogientest (FAT 7 - 8) von BELSER, ANGER und BARGMANN (1965).

(9) Figurengruppentest

Dieser Test ist wiederum identisch mit dem gleichnamigen, 1964 verwendeten Verfahren. Die Bearbeitungszeit wurde jedoch reduziert.

Dieser und der Wortanalogietest sollen zur besseren Abschätzung der Fähigkeit des schlußfolgernden Denkens beitragen.

(10) Textaufgabentest

Dieser Test besteht aus sogenannten "eingekleideten" Rechenaufgaben, die zu lösen sind. Dabei dürfen Nebenrechnungen auf einem zusätzlichen Blatt ausgeführt werden.

Aufgabenbeispiel:

Ein Kleinwagen verbraucht auf 14 km
1 Liter Benzin. Wie viele km kann er
mit 15 Litern fahren? _____ km

Anzahl der Items: 24

(11) Diktattest

Das Diktat "Sind Atomraketen möglich?" aus HOPF, THIEL,
REININGHAUS (s. J.) besteht aus 126 Wörtern, die nach
dem Diktieren fehlerfrei zu schreiben sind. Die fehler-
frei geschriebenen Wörter wurden addiert und als Kenn-
wert verwendet.

Dieser und der Textaufgabentest sollen als schulleistungs-
nahe Verfahren Auskunft geben über den Zusammenhang von
Testleistungen mit den Schulnoten Deutsch und Rechnen/
Mathematik.

Fragebögen

(12) ENRA - Fragebogen

Der Fragebogen enthält 12 Fragen zu "Extraversion" (E),
25 Fragen zu "Neurotizismus" (N), 15 Fragen zu "Rigidität"
(R), und 12 Fragen zu "Aufrichtigkeit" (A). Herkunft:
TODT.

Beispiel:

Fühlst du dich bei neuen Aufgaben unsicher? ja nein
Das am ehesten Zutreffende muß angestrichen werden. Die
Fragen zu den einzelnen Persönlichkeitsmerkmalen wurden
zu jeweils einer Punktsomme addiert.

(13) I-A-R-Fragebogen zur internalen Kontrolle
Der Fragebogen wurden von CRANDALL et al. (1965) entwickelt und von MEYER (196) Übersetzt und - wo erforderlich - geändert (vgl. WINDHEUSER 1968). Mit den Fragen sollen Aussagen über die interne Kontrolle der Probanden in intellektuell-schulischen Erfolgs- oder Mißerfolgssituationen (Intellectuell Achievement Responsibility, I-A-R) möglich sein.

Die Antworten auf die Fragen wurden je Proband zu zwei Punktsommen addiert; je 15 Fragen beziehen sich auf die Komplexe interne Kontrolle bei Erfolg (IE) bzw. Mißerfolg (IM). Außerdem wurden mit Hilfe dieser Punktsommen zwei weitere Kennwerte gebildet: Ein Gesamtwert der internalen Kontrolle, $IG = IE + IM$, und ein Nettowert, $IN = (IE - IM) + 20$, analog dem Nettowert, den HECKHAUSEN (1963) zur Bestimmung der Leistungsmotivation verwendet hat. (Der konstante Summand dient der Vermeidung negativer IN.)

Beispielfrage:

Woran liegt es, wenn du in der Schule irgendetwas schnell lernst?

- A ☐ Weil du gut aufgepaßt hast?
B ☐ Weil der Lehrer es gut erklärt hat?

Der Kreis vor dem zutreffenden Statement ist anzukreuzen.

(14) Interessen- und Milieufragebogen (I-M-F)

Dieser Fragebogen wurde in zwei Versionen entwickelt:
Eine für Hauptschüler und eine für Real- und Oberschüler,
um so zum Bereich der Schule, des Elternhauses und der

Absichten der Schüler nach dem Verlassen der Schule zusätzliche Informationen zu erhalten, die die Bedingungen von Erfolg und Versagen besser abzuklären gestatten sollten. (Den Sonderschülern wurde dieser Fragebogen nicht vorgelegt.)

Beispielfrage:

Hast du ein Zimmer, in dem du gewöhnlich allein bist, wenn du Hausaufgaben machst?

immer fast immer meistens selten nie

o-----o-----o-----o-----o

Anzahl der Items: Version VS 35; Version RS/OS 34.

Außer Fragen mit Mehrfachwahlantworten waren auch solche mit disjunktiven Antworten und auch offene Fragen (ohne Antwortvorgaben) im Fragebogen enthalten.

(16) Lehrerfragebogen zum Leistungs- und Sozialverhalten und zum voraussichtlichen bzw. möglichen Schulabschluß der Schüler im allgemeinbildenden Schulwesen

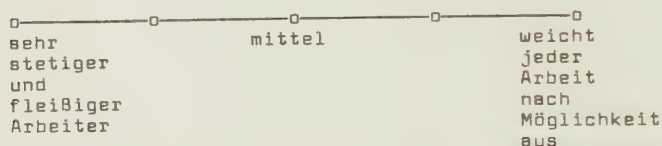
Dieser, den Lehrern vorgelegte Fragebogen, ist wiederum analog denen konstruiert worden, die den Lehrern 1960/61 und 1964 vorgelegt worden waren. Dabei war natürlich das Alter der Schüler von nunmehr 13 - 14 Jahren zu berücksichtigen.

Der Fragebogen wurde in drei Versionen entwickelt; je eine für Haupt-, Real- und Oberschüler. Die Unterschiede in den einzelnen Versionen bestanden jedoch lediglich in den Fragen, die - nach Meinung der Lehrer - den wahrscheinlichen Schulabschluß der Schüler betrafen; bei den Hauptschülern war nach einer Prognose für den erfolgreichen Abschluß einer Aufbaurealschule nach Abschluß

der Hauptschule gefragt.

Beispielfrage:

Welchen Eindruck haben Sie vom Arbeitsverhalten des Kindes?



Anzahl der Items, Version HS: 18

Anzahl der Items, Version RS: 19

Anzahl der Items, Version OS: 19

Nach faktorenanalytischen Berechnungen wurden die Antworten einiger Fragen zu Punktwertgruppen addiert.

Außerdem wurden zwei Gesamtkore je Pb gebildet.

Tabelle 6 : Zusammensetzung der Punktwertgruppen

Punktwertgruppe	lfd. Nr. der Fragen
Kontakt	3; 4
Leistung	5; 6; 7; 8; 9; 10
Entwicklung	11; 12
Milieu I	14; 16
Milieu II	14; 15; 16
Gesamtkore I	3 - 12; 14 - 16
Gesamtkore II	3 - 16 (Frage 13 umgepolt)

(16) Elternfragebogen (E-M-F)

Dieser von ZINGELER entwickelte Fragebogen wurde - incl. Freiumschlag - den Eltern der Schüler mit der Bitte um Bearbeitung und Rücksendung zugeschickt. Er enthält Fragen zur Größe und Vollständigkeit der Familie, zu den Wohnverhältnissen, dem Familieneinkommen, zur Schulbildung der Eltern und zu deren Berufstätigkeit bzw. dem Beruf des Vaters.

Beispielfrage:

Wohnen Sie zur Miete oder in einer Eigentumswohnung
oder im eigenen Haus?

Mietwohnung	
Eigentumswohnung	
eigenes Haus	

Anzahl der Items: 19

.3.4

Schulleistungsvariablen

Außer den psychologischen Verfahren wurden für die Jahre 1961 - 1968 von den jeweils erreichbaren Schülern unserer Stichprobe die Versetzungsnoten in Deutsch und Rechnen/ Mathematik, von den Real- und Oberschülern zusätzlich noch die Versetzungsnoten in den Fremdsprachen erhoben. Ferner haben wir über den genannten Zeitraum auch Angaben über die Versetzung oder Nicht-Versetzung der Schüler eingeholt. Für die Oberschüler reichen diese Informationen bis 1974 (vgl. FIEGENBAUM 1974).

Untersuchung Frühjahr 1972 (nur Oberschüler)

Diese Untersuchung wurde von FIEGENBAUM (1974) durchgeführt. In dieser Arbeit sollte versucht werden, den schulischen Werdegang aller jemals auf dem direkten Weg (GS → OS) eine Oberschule (= Gymnasium) besuchenden Schüler der Ausgangsstichprobe (N = 315) bis zum Abitur resp. dem (vorzeitigen) Verlassen der Oberschule zu verfolgen und - wenn möglich - mit Hilfe einiger psychologischer Variablen Erklärungsmöglichkeiten hinsichtlich des Schulschicksals Einzelner oder Gruppen zu erhalten.

Die Substichprobe umfaßt 71 Schüler (von 79, s. Tabelle 2), davon 24 Mädchen. Von diesen Oberschülern/innen wurden 46 in 6 Gruppen von 6 - 12 Pbn mit dem Intelligenz-Struktur-Test (IST 70) von AMTHAUER (1970) untersucht. An 2 weiteren Untersuchungen nahmen jeweils nur 2 bzw. 1 Schüler teil. Außer dem IST 70 wurde von den Pbn auch ein von FIEGENBAUM für diesen Zweck entwickelter Fragebogen zur Schulsituation und den Berufs- und Studieninteressen bearbeitet. Die Datenerhebungen fanden in der Zeit vom 14. 04. - 25. 05. 1972 statt, bei den zuerst genannten 6 Gruppen jeweils nachmittags im Psychologischen Institut Münster. Um den Gruppencharakter bei der Durchführung zu erhalten, wurden in zwei Fällen die Gruppen mit weiteren Pbn aufgefüllt, deren Tests und Fragebögen jedoch nicht in die Analyse einbezogen wurden. FIEGENBAUM hat in seiner Arbeit nur

die Tests und Fragebögen der 40 Schüler berücksichtigt, die 1972 noch eine Oberschule besuchten. Der Schulstatus, d. h. wann und mit welchem Abschluß die Jugendlichen die Schule verlassen haben wurde jedoch für 71 Schüler ermittelt.

5.0 DURCHFÜHRUNG DER UNTERSUCHUNGEN

5.1 Ausgangsuntersuchung Ostern 1960/1961

Wie bereits bei der Stichprobenbeschreibung ausgeführt, haben wir es hier mit 2 Primärstichproben zu tun, den altersgemäß Eingeschulten und den Zurückgestellten. Demgemäß seien auch die Durchführungen der Untersuchungen für die beiden Stichproben getrennt dargestellt.

Die altersgemäß Eingeschulten:

Die 261 Schüler dieses Kollektivs verteilten sich auf die erwähnten 5 Schulen (6 Klassen) in und um Münster und wurden von 4 Studenten der Psychologie in der Zeit vom 25. 04. 1960 - 16. 05. 1960 (Stichtag: 30.04. 1960) in 19 Gruppen von je 10 - 18 Kindern, im Mittel 14 Kinder, untersucht. Die Schüler/innen waren am Stichtag der Untersuchung zwischen 5;10 - 6;9 Jahre alt.

Bei allen Gruppen wurde zuerst der Grundleistungstest, bei 9 Gruppen anschließend der Auslesetest für Schulneulinge und dann der Göppinger Schulreifetest erhoben. Bei 10 Gruppen wurde die Reihenfolge der beiden zuletzt genannten Tests vertauscht. "Mit der wechselweisen Anwendung der beiden großen Schulreifetests wollten wir erreichen, daß ein bei der jeweils ersten Prüfung zu erwartender Übungsgewinn sich gleichmäßig auf die Gesamtergebnisse der beiden Tests auswirken sollte" (KIEKHEBEN 1960, S. 16).

Die Durchführung erfolgte bei allen Gruppen an zwei aufeinander folgenden Tagen, wobei der für die einzelnen Gruppen zuletzt genannte Test am zweiten Tag durchgeführt wurde.

Der Fragebogen zur seelisch-sozialen Schulreife wurde den Lehrerinnen ca. 8 Wochen nach Schulbeginn der Kinder zur Beantwortung übergeben und ca. eine Woche später wieder eingesammelt (STROTHMANN 1960).

Mit allen Schülern, die nach den Schulreifetests als "nicht-schulreif" oder "fraglich schulreif" klassifiziert werden mußten ($N = 40$), wurde in den Wochen nach Auswertung der Schulreifetests noch der Intelligenztest von SIMON-BOBERTAG in der Bearbeitung von I. NORDEN (1953) durchgeführt. Als Kontrollgruppe wurde eine Gruppe von $N = 10$ schulreifen Kindern hinzugenommen (REMBERG 1960).

Die Zurückgestellten: Diese Gruppe bestand 1961 aus insgesamt 74 Kindern. Vierzehn davon brauchten jedoch nicht untersucht zu werden, da sie schon 1960 an der Untersuchung teilgenommen hatten, 9 von ihnen jedoch nachträglich zurückgestellt worden waren und 5 die Versetzung in die 2. Klasse nicht erlangt hatten.

Die Schüler gehörten - wie erwähnt - zum Einzugsbereich unserer 5 Münsteraner Schulen. Die Untersuchungen erfolgten in der 2. Aprilhälfte in 8 Gruppen zu 12 - 14 Kindern. Einige Gruppen in den Stadtrandschulen enthielten nur 5 Schüler.

Auch hier wurde immer zuerst der Grundleistungstest und dann der Auslesetest für Schulneulinge erhoben. Auf die Durchführung des Göppinger Schulreifetests wurde verzichtet.

Nicht verzichtet wurde auf die Beantwortung des Fragebogens zur seelisch-sozialen Schulreife und die Einzeluntersuchung der Kinder mit dem Intelligenztest nach SIMON-BOBERTAG. Diese Erhebung wurde bei allen Zurückgestellten vorgenommen (HOJAK 1962).

Das Alter der Zurückgestellten lag zum Stichtag der Untersuchung 1961 zwischen 6;10 - 7;9 Jahren. Das höhere Alter - im Vergleich zu den altersgemäß Einschulten - muß bei der Bewertung der Ergebnisse besonders beachtet werden.

5.2 Untersuchung Frühjahr 1964

Da bei dieser Untersuchung die Schüler nicht mehr nach dem Einschulungstermin unterteilt wurden, entfällt eine entsprechend getrennte Untersuchungsbeschreibung.

Die Ausgangsstichprobe hatte sich aus verschiedenen Gründen von 341 auf 251 (74%) reduziert (KEMMLER 1967, S. 31). Die Schüler/innen verteilten sich mittlerweile auf 6 Grund- und 2 Sonderschulen und befanden sich im 2., 3. oder 4. Schuljahr der Grundschule bzw. im 1. oder 2. Schuljahr der Sonderschule.

Die Untersuchungen wurden von 5 Studenten der Psychologie durchgeführt, die je einen fest umrissenen Aufgabenbereich

hatten. Der Untersuchungszeitraum erstreckte sich vom 08. 01. 1964 - 26. 02. 1964. Als Stichtag wurde der 31. 01. 1964 festgelegt. Die Schüler/innen waren zu diesem Zeitpunkt zwischen 9;7 - 10;6 Jahre alt. Die Testdurchführungen wurden in Gruppen von 9 - 16 Schülern vorgenommen, bei den Sonderschülern betrug die Gruppengröße 4 - 6 Schüler, wobei es notwendig wurde, altersgleiche, jedoch nicht zur Längsschnittstichprobe gehörende Schüler zu einigen Untersuchungen hinzuzunehmen, um den Gruppencharakter der Testerhebungen zu gewährleisten. Die Testergebnisse solcher "Auffüllkinder" blieben unausgewertet.

Die einzelnen Verfahren wurden bei allen Gruppen in der gleichen Reihenfolge erhoben.

Die Reihenfolge der Tests war:

("standardisierte" Einführung)

Bilderwortschatztest
Wortschatztest
Raumerfassungstest
Wortgruppentest
Figurengruppentest
Wahrnehmungsschnelligkeitstest
Zahlentest
Kinderfragebogen
Wissenstest
Wörtertrennentest
Wörtersuchen-Test
Fehlersuchen-Test
Baumtest (Die Aufgabe der Schüler bestand darin, einen

schönen Obstbaum zu malen. Daran sollte immer dann gearbeitet werden, wenn jemand mit einem der übrigen Tests vorzeitig fertig war. Auf diese Weise sollten gegenseitige Störungen minimalisiert werden. Am Ende der Untersuchung wurde der Obstbaum von allen Schülern fertiggestellt. Der Baumtest blieb jedoch unausgewertet.)

("standardisierte" Abschlusserklärung über die Untersuchungsabsicht)

Diese Reihenfolge war, da die Erhebungen während der üblichen Unterrichtszeit vorgenommen wurden, notwendig, um z. B. die den Schülern gewohnten Pausen einhalten zu können (ZINGELER 1965), schließt aber Positioneffekte bei bestimmten Tests nicht aus (vgl. LUCHINS 1942; HERBIG 1972). (Positioneffekte werden bei Testbatterien, die in der Regel in ihrer Abfolge festgelegt sind bzw. aus Gründen der Standarddurchführungsbedingungen festgelegt sein müssen, in Kauf genommen.)

Auf einen dieser möglichen Effekte soll hier nachdrücklich hingewiesen werden, da er nachträglich als Versuchsfehler klassifiziert werden muß und deutlich macht, welch große Sorgfalt der Testfolge und dem Instruktionsverständnis gewidmet werden muß:

ERTEL hatte einige Jahre nach dieser Untersuchung bei einer eigenen ebenfalls die Primary Mental Abilities (vgl. KEMMLER & LANGHEINRICH 1967) verwendet. Beim Test Wortgruppen fragte ihn eine Schülerin, ob denn auch das erste Wort einer Aufgabenreihe als richtige Lösung in Frage komme. Nach dieser Mitteilung ERTEL's haben wir alle Testhefte daraufhin durchgesehen und gefunden, daß ca. 7% (von 251) der Probanden unserer Längsschnittuntersuchung nie das erste Wort einer Aufgabenzeile als Lösung angegeben hatten. Daraufhin haben wir auch die 472 entsprechenden Testhefte der Querschnittuntersuchung von KEMMLER (1967) durchgesehen. Diese Stichprobe setzte sich nach den Schulleistungen aus den jeweils 2 besten Schülern/innen einer Klasse, insgesamt N=132, und den jeweils 5 - 6 schwächsten dieser Klassen, insgesamt N=340, zusammen. Von den 132 Besten hatten ca. 5%, von den 340 Schwächsten ca. 9% nie das erste Wort einer Zeile als Lösung angegeben.

Als Beispiel sei hier die 4. Originalaufgabe des Wortgruppentests wiedergegeben.

Biene	Rose	Nelke	Tulpe
-------	------	-------	-------

Da die Anweisung lautete, das nicht in die Reihe passende Wort durchzustreichen, wäre hier das Wort "Biene" zu markieren gewesen.

Welche Erklärungen lassen sich nun für das Mißverständnis finden?

1. Von den drei Beispielaufgaben dieses Tests war einmal das dritte Wort und zweimal das zweite die richtige Lösung, d. h. die Anordnung der richtigen Lösungen bei den Übungsaufgaben war nicht optimal und kann zu Verwirrungen geführt haben.

2. Dem Wortgruppentest ist immer der Raumerfassungstest vorausgegangen. Beim Raumerfassungstest bildet jeweils die erste Figur einer Aufgabenreihe den Standard- oder Bezugsreiz, d. h. die Pbn müssen in einer Reihe jeweils die Figur ankreuzen, die zusammen mit dem Standardreiz ein Quadrat ergibt. Es wäre denkbar, daß die Reihenfolge der Durchführung der Subtests hier einen Positionseffekt in Form eines negativen Lösungsstrategietransfers verursacht hat.

Mag sein, daß aber auch ganz andere kognitive Prozesse bei den betroffenen Probanden das generelle Fehlverhalten ausgelöst haben. Wir wollten jedenfalls auf die Schilderung dieses Versuchsfehlers nicht verzichten, weil der Mitteilung von Versuchsanordnungen, Durchführungsschwierigkeiten und -fehlern etc. in der Literatur fast immer zu wenig Raum gegeben wird bzw. unsere Unzulänglichkeiten kaum eine Ausnahmeerscheinung sein dürften, die vernachlässigt werden dürfe.

5.3

Untersuchung Frühjahr 1968

In diese Untersuchung konnten noch 224 Pbn (66% von 341) einbezogen werden. Die Schüler/innen verteilten sich auf insgesamt 34 Schulen (s. Stichprobenbeschreibung).

Die Untersuchungen (in Gruppen) wurden von der Projektleiterin, einem Diplom-Psychologen und 6 Studenten/innen der Psychologie durchgeführt. Dabei übernahm ein Student neben dem Diplom-Psychologen die Rolle eines zweiten Versuchsleiters, während den übrigen Studenten/innen wieder jeweils fest umrissene Versuchshelfertätigkeiten übertragen wurden (vgl. STURZEBECKER 1968; REICHEL 1969). Der Untersuchungszeitraum erstreckte sich vom 22. 01. 1968 - 28. 03. 1968. Als Stichtag wurde der 15. 02. 1968 festgelegt. Zu diesem Zeitpunkt waren die Schüler zwischen 13;6 - 14;5 Jahre alt. Die Testdurchführungen erfolgten in 24 Gruppen zu je 10 - 14 Pbn. Infolge der Verteilung der Schüler über 34 Schulen wurde es bei 12 Gruppen notwendig, einige alters- und schulstatusgleiche Schüler

aus den infrage kommenden Klassen hinzuzunehmen, um die gewünschte mittlere Gruppengröße zu erhalten. Die Ergebnisse der insgesamt 64 "Auffüllschüler" blieben unausgewertet.

In zwei weiteren Gruppen zu 8 bzw. 4 Schülern wurden die Sonderschüler untersucht. Diese Testdurchführungen kamen beinahe Einzelerhebungen gleich. So bereitete es den Sonderschülern nicht selten Schwierigkeiten, die Testanweisung zu verstehen, oder sie arbeiteten während der Zeit, die für die Bearbeitung eines Tests zur Verfügung stand, nicht voll durch. Für die Durchführung aller Tests einschl. Fragebögen waren bei den Sonderschülern zwei Vormittage (mit zwei Versuchshelfern) nötig. Bei allen übrigen Gruppen genügte ein Vormittag.

Die Schwierigkeiten bei der Untersuchung der Sonderschüler bedeutete eine Verletzung der gewünschten Standardbedingungen, was bei der Darstellung und Erörterung der Ergebnisse besonders zu werten sein wird.

Die Reihenfolge der Tests war stets:

("standardisierte" Einführung)

Wortschatztest
Figurenvergleichstest
Wortanalogietest
Textaufgabentest
Wahrnehmungsgeschwindigkeitstest
ENRA-Fragebogen
I-A-R-Fragebogen
Milieufragebogen
Buchstabenreihentest
Figurengruppentest
Wortgruppentest
Diktattest

Konzentrationsleistungstest (KLT), die Leistungen der ersten 7 Minuten wurden als Zahlentest gewertet.

("standardisierte" Abschlußerklärung über die Untersuchungsabsicht)

Auch bei dieser Untersuchung wurde die Testfolge so gewählt, daß möglichst die Unterrichtspausen eingehalten werden konnten. Bezüglich möglicher Positionseffekte gilt Ähnliches wie bereits erwähnt.

Weitere Details zu Durchführungen dieser Untersuchung sind in einem Forschungsbericht von KEMMLER (1968) enthalten.

DIE FORMALE AUSWERTUNG DER TEST- UND FRAGEBOGENER- HEBUNGEN

Die Auswertung der Test- und Fragebogenergebnisse erfolgte in der Regel von den an der jeweiligen Untersuchung beteiligten Studenten. In einigen Fällen wurden noch zusätzliche Studenten hinzugezogen.

Zur Auswertung der verwendeten psychologischen Verfahren wollen wir zunächst erwähnen, auf welche Art und Weise die Probanden zu ihren Ergebnissen gelangen konnten, weil so gleichzeitig Aufschluß über die Auswerterobjektivität zu erlangen ist:

1. Der größte Teil aller Testaufgaben war so gestaltet, daß die richtige Lösung jeweils unter mehreren Alternativlösungen "multiple choice" enthalten war, die von den Probanden angekreuzt werden mußte.
2. Bei den Zahlen- bzw. Rechentests war die jeweilige Lösung der einzelnen Aufgaben zu ermitteln und anzugeben.
3. Auch bei den Tests Wörtersuchen und Wörtertrennen war die richtige Lösung von den Pbn jeweils eindeutig zu ermitteln. Dies gilt auch für den Diktat-test, für dessen Auswertung ein Bewertungsschema erstellt worden war (s. Beschreibung der Tests).

4. Die den Schülern, Eltern und Lehrern vorgelegten Fragebögen waren entweder mit vorformulierten Alternativantworten zu den einzelnen Fragen versehen oder es waren quantitative Angaben erbeten worden, die von den Auswertern wiederum eindeutig zu klassifizieren waren. Eine Ausnahme bildete die Frage über dessen berufliche Haupttätigkeit. Zur Auswertung dieser Fragen wurde von ZINGELER auf Grund der erhaltenen Angaben ein Klassifikationsschema erstellt (MAST 1969).

5. Am schwierigsten war - infolge der Strukturierung der Tests - die Auswertung einiger Aufgaben bzw. Subtests unter den Schulreifeverfahren. Beispielshaft sei hier der Grundleistungstest erwähnt, der in der Längsschnittuntersuchung in der verkürzten Fassung von drei Aufgaben verwendet wurde. Die Auswertung der Aufgaben führte bei den Auswertern im Mittelbereich der Bewertungsmöglichkeiten am ehesten zu Differenzen. KEMMLER und HECKHAUSEN (1962) haben für diesen Test ein eigenes Bewertungsschema entwickelt.

Wo möglich, wurden für die Auswertung Schablonen angefertigt. Dessen ungeachtet wurden die Aufgaben zweimal - von verschiedenen Auswertern - ausgewertet, um auch simple Lesefehler, die durchaus vorgekommen sind, auszuschließen.

Die Summe der richtigen Lösungen wurde bei den Leistungstests jeweils als Punktrohwerte verwendet. Die Rohwerte aus den multiple choice Tests der Primary Mental Abilities wurden mit Ausnahme des Figurenvergleichstests zufallskorrigiert (LIENERT 1967). Dieses Vorgehen erscheint zumindest für die PMA 9-10 nachträglich gerechtfertigt, wie die Kontrolle der Resultate des Wortgruppentests anlässlich der Mitteilung ERTEL's gezeigt hat. Aus Gründen der besseren Vergleichbarkeit der Ergebnisse ist die Zufallskorrektur auch bei den PMA 13-14 durchgeführt worden. +)

6. Die Auswertung des IST 70 war problemlos, da sie mittels Schablonen und die Ermittlung der Standardwerte nach den vorgegebenen Normtabellen erfolgte. Um eine möglichst hohe Objektivität zu erzielen, wurde die Auswertung des Fragebogens von FIEGENBAUM zur Schulsituation und den Berufs- und Studieninteressen der Oberschüler nach einem vorgegebenen Auswertungsplan von zwei Auswertern unabhängig voneinander vorgenommen. Differenzen wurden anschließend von den Auswertern gemeinsam geklärt (vgl. FIEGENBAUM 1974).

+) Detailliertere Angaben zur Auswertung finden sich in Berichten von Lilly KEMMLER für die Deutsche Forschungsgemeinschaft. Die Berichte können im Psychologischen Institut Münster eingesehen werden.

7.0

DIE INSTRUMENTELLEN EIGENSCHAFTEN DER VERFAHREN

In diesem Kapitel sollen Angaben über die verwendeten Verfahren als Meßinstrumente gemacht werden, d.h. darüber, inwieweit die Tests die Anforderungen erfüllen, die an sie als diagnostische und/oder prognostische Hilfsmittel zu stellen sind. Dabei sei auch auf die bereits bei KEMMLER (KEMMLER 1967, S. 33-40; 1967a und 1976) aufgeführten Angaben verwiesen. Unser Hauptinteresse gilt den PMA 9-10 und PMA 13-14, daneben auch den Schulreifetests. Die zentralen Merkmale zur Beurteilung der Brauchbarkeit von Tests als Meßverfahren lassen sich durch Informationen zur Objektivität, Reliabilität und Validität darstellen. Damit im Zusammenhang stehen Fragen zur Rohwertverteilung der Testwerte, der Teststandardisierung, der Skorenbildung und der Standardmeßfehler.

7.1

PRÜFUNG AUF NORMALVERTEILUNG

a) Die beiden Schulreifetests (Untersuchung 1960)

Auf eine differenzierte Darstellung von Kennwerten zu diesen Tests wird hier verzichtet, denn, da diese Tests als Schulreifetests und nicht als Tests zur Feststellung allgemeiner Intelligenz - wie die PMA -

konstruiert wurden, können im Hinblick auf die hier primär interessierende Intelligenz der Schüler auch gar keine besonderen Erwartungen an die Ergebnisse aus den Schulreifetests gestellt werden. Zudem dürften bereits den beiden Tests verschiedene intentionale Gehalte des Begriffs "Schulreife" zugrunde liegen. Von KIEKHEDEN (1960) wurden für diese Tests Prüfungen auf Normalverteilungen vorgenommen. Dabei ergab sich eine Normalverteilung der Gesamtrohwerte des Auslesetests für Schulneulinge, Form B ("Münchener") und eine J-Verteilung der Gesamtrohwerte des Grundleistungstests ("KERN").

b) Die sieben PMA 9-10 (Untersuchung 1964)

Die Verteilung der Testwerte der Untertests dieser Testbatterie wurde bereits von LANGNER (1965) überprüft (vgl. KEMMLER 1967, S. 34). Dazu wurden die Schüler nach Altershalbjahren eingeteilt, weil die Hypothese bestand, in diesem Alter wachse die kognitive Leistungsfähigkeit. Die gleiche Alters-einteilung wurde damals bei den auffindbaren Schülern dieser Längsschnittuntersuchung (N = 251) vorgenommen. Die Verteilungsprüfungen waren im Ergebnis gleich: Eine signifikante Abweichung von der Normalverteilung ergab sich nur im Zahlentest für die

Gruppe der jüngeren Schüler. Die Verteilung der Rohwerte dieser Gruppe ist mehr- bzw. breitgipflich.

Um nun für die vorliegende Arbeit auch über die Rohwertverteilungen zusätzliche Hinweise hinsichtlich einer möglicherweise überzufälligen Schrumpfung der verbliebenen Stichprobe zu erhalten, wurden erneute Normalverteilungsprüfungen vorgenommen. Aus Vergleichsgründen wurden die Pbn in 2 Alterhalbjahre geteilt. So entstanden die Gruppe der "Jüngere" mit $N = 126$ und die der "Ältere" mit $N = 98$ Schülern. MEILI & STEINGRÜBER (1978, S. 283) halten für eine Verteilungsprüfung eine untere Probandenzahl von $N = 100$ für erforderlich. Diese Zahl wird von den älteren Schülern knapp unterschritten. Von MITTENECKER (1964, S. 8) werden etwa 6-12 Rohwertklassen gefordert. LIENERT (1967, S. 173) und MEILI (1961, S. 423) fordern besonders bei kleinen Analysestichproben, deren Meßwerte noch dazu über einen weiten Bereich streuen, die Testpunkte in 10-15 Klassen zusammenzufassen, um so auffällige Irregularitäten auszugleichen. MEILI & STEINGRÜBER (1978, S. 283) fordern eine Zusammenfassung der Testwerte zu 10-20 Klassen. Im vorliegenden Falle schwanken die Klassenzahlen zwischen 7-12. Die Normalverteilungsprüfung selbst

wurde nach LIENERT (1967, S. 171-175) vorgenommen, mit Hilfe eines EDV-Programms von WINDHEUSER ¹⁾. Dabei werden die Verteilungen hinsichtlich ihrer "Güte der Anpassung", ihrer "Schiefe" und ihres "Exzesses" gerprüft. Abweichungen werden dann als bedeutsam angesehen, wenn mindestens einer der drei Kennwerte auf dem 1%-Niveau signifikant ist (LIENERT 1967, S. 171ff). WINDHEUSER's Programm bildet - vereinfacht dargestellt - maximal 15 Rohwertklassen und überprüft die sich daraus ergebenden Kennwerte auf Signifikanz. Wird ein oder mehrere Kennwerte auf dem 1%-Niveau signifikant, so wird durch Vergrößerung der Klassenintervalle und Verringerung der Zahl der Rohwertklassen die Verteilungsbreite schrittweise reduziert, wobei die statistischen Kennwerte jeweils neu ermittelt und auf Signifikanz überprüft werden. Bei einer Mindestklassenzahl wird das Verfahren abgebrochen.

Die Kennwerte, die sich auf diese Weise ergeben haben, sind in Tabelle 7 zusammengefaßt. Gegenüber den oben zitierten früheren Ergebnissen, weist der Zalentest hier keine signifikanten (1%) Verteilungsabweichungen auf. Dagegen ist die Rohwert-

¹⁾ Dem Kollegen Jochen WINDHEUSER sei für die Überlassung seines EDV-Programms herzlich gedankt.

Tabelle 7 :

Prüfung der Teststreuwerte auf Normalverteilung,
getrennt nach zwei Altershalbjahren
Jüngeres Halbjahr N = 126, Älteres Halbjahr N = 98

Test	Alter	Güte- chi ²	Anpassung df	Schief e S	Schief e (S) t-Wert	E	Exzeß (E) t-Wert
(1) Bilderwortschatz	Jüngere	5,45	6	0,09	0,39	2,39	-1,39
	Ältere	12,54	6	0,05	0,21	2,48	-1,05
(2) Wortschatz	Jüngere	12,06	9	-0,33	-1,53	2,46	-1,24
	Ältere	5,60	7	-0,44	-1,78	2,62	-0,77
(3) Raumerfassung	Jüngere	5,27	6	-0,04	-0,18	2,74	-0,60
	Ältere	3,42	4	0,29	1,17	2,60	-0,82
(4) Figurengruppen	Jüngere	5,78	6	-0,38	-1,72	2,78	-0,50
	Ältere	5,98	4	0,10	0,42	2,55	-0,91
(5) Wortgruppen	Jüngere	10,90 +	4	-0,60 ++	-2,74	2,95	-0,11
	Ältere	16,26 ++	4	-0,60 +	-2,44	3,03	0,05
(6) Wahrnehmungss.	Jüngere	10,71	6	0,04	0,20	2,74	-0,59
	Ältere	3,48	4	0,14	0,57	3,10	0,20
(7) Zahlen	Jüngere	10,10	6	-0,44 +	-2,01	2,59	-0,94
	Ältere	7,14	6	-0,41	-1,67	3,64	1,29
(8) Wortschatz	Jüngere	10,42	8	-0,01	-0,05	2,29	-1,62
	Ältere	10,20	7	-0,04	-0,18	2,52	-0,98
(9) Figurenvergleich	Jüngere	12,94	7	-0,29	-1,33	2,23	-1,76
	Ältere	10,05	5	-0,13	-0,54	1,31 +	-2,21
(10) Buchstabenreihen	Jüngere	13,12 +	6	-0,31	-1,43	2,07 +	-2,13
	Ältere	7,11	5	-0,26	1,04	2,72	-0,56
(11) Wortgruppen	Jüngere	5,93	8	-0,07	-0,30	2,56	-1,01
	Ältere	8,79	8	-0,04	-0,14	2,49	-1,32
(12) Wahrnehmungssch.	Jüngere	5,67	5	0,31	-1,43	3,13	0,30
	Ältere	4,08	5	0,48	-1,95	3,05	0,11
(13) Zahlen	Jüngere	1,90	4	-0,02 +	-0,09	2,56	-1,00
	Ältere	0,93	5	0,50	2,04	2,86	-0,26

+ entspricht einem Signifikanzniveau von 5 %, ++ einem von 1 %

Bei positivem t-Wert ist die Verteilung linkschief bzw. der Exzeß schmaligipflig, bei negativem t-Wert ist die Verteilung rechtsschief bzw. der Exzeß breitigipflig.

verteilung des Tests Wortgruppen für die Gruppe der älteren Probanden mehrgipflich ($p < 1\%$), mit deutlicher Tendenz ($p < 5\%$) zur Rechtsasymmetrie. Für die Gruppe der jüngeren Probanden verschärft sich dieser Befund für die Schiefe noch weiter in Richtung der Rechtsasymmetrie ($p < 1\%$), wogegen die Mehrgipflichkeit an Deutlichkeit zurückgeht ($p < 5\%$). Es kann vermutet werden, daß diese Resultate im Zusammenhang mit dem geraume Zeit nach der Testerhebung aufgedeckten Mangel an Instruktionsverständnis bei ca. 7% der Probanden eine Ursache haben (vgl. S. 86).

c) Die sechs PMA 13-14 (Untersuchung 1968)

Die Verteilungsprüfungen wurden in genau der gleichen Weise vorgenommen wie bei den PMA 9-10. Auch die statistischen Kennwerte dieser Tests sind in Tabelle 7 enthalten. Wie die Ergebnisse zeigen, haben sich für keinen Test und für keine Altersgruppe Werte ergeben, die die Aufgabe der Normalverteilungshypothese erfordern würden.

Im Gegensatz zu den Ergebnissen über die beiden Altershalbjahre steht der Befund, den die ebenfalls durchgeführten Normalverteilungsprüfungen der Tests über die Gesamtstichprobe ($N = 224$) für die Tests Figurenvergleich und Buchstabenreihen ergeben haben: Für den Test Figurenvergleich ergab sich hinsichtlich der Modalität eine auf dem 0,1 %-Niveau signifikant abweichende Verteilung und eine deutliche Tendenz ($p < 5\%$) zur Breitgipflichkeit. Beim Test Buchstabenreihen ist der Prüfwert für die Modalität der Verteilung auf dem 1 %-Niveau signifikant und die Schiefe weist eine deutliche Tendenz ($p < 5\%$) zur Rechtsasymmetrie auf, d.h., für die Tests Figurenvergleich und Buchstabenreihen muß die Normalverteilungshypothese der Testwerte - bezogen auf die Gesamtstichprobe - verworfen werden.

Es ist sicher nicht ganz einfach, (zutreffende) Erklärungen für diese Abweichungen zu finden. KEMMLER (1976, S. 38) hält es für wahrscheinlich, daß die Verlängerung der Bearbeitungszeit von 5 auf 8 Min. beim Test Figurenvergleich (gegenüber der von THURSTONE vorgegebenen Zeit) die Wahrscheinlichkeit, mehr richtige Lösungen zu erzielen, erhöht und damit die nicht-normale Verteilung verursacht habe. Dieses Argument ist sicher nicht zu widerlegen, enthielte sein volles Gewicht aber erst dann, wenn die

Zeitverlängerung auch eine deutliche Rechtsasymmetrie der Verteilung - auch bei den Substichproben - aufwiese, dies um so mehr, als die maximal mögliche Zahl richtiger Lösungen sowohl von Schülern der HS,RS und OS erreicht wurde (vgl. Tabelle 8). Eine deutliche Rechtsasymmetrie liegt jedoch nicht vor. Die Normalverteilungsabweichungen müssen daher noch weitere Gründe haben, die in der Testkonstruktion selbst, aber auch im möglicherweise unterschiedlichen Aufforderungscharakter der Aufgaben für die Schüler der verschiedenen Schularten liegen können. Aber auch die Genese sowie die Größe der Substichproben müssen als mögliche verursachende Faktoren in Betracht gezogen werden. Die hier zusätzlich genannten Verursachungsmöglichkeiten könnten - bezogen auf die Gesamtstichprobe - auch für die nicht-normale Rohwertverteilung des Tests Buchstabenreihen zutreffen. Eine genauere Analyse der Verursachungsbedingungen erwartungswidriger Verteilungen müßte u.E. auch eine Besprechung der Testergebnisse mit den Probanden einschließen, was hier nachträglich jedoch nicht mehr möglich ist.

Die Tabellen 9a und 9b mit den nach Schulart getrennten Rohwertverteilungen machen deutlich, daß die Mehrgipfligkeit der Verteilungen beim Test

Tabelle 8 : Mittelwerte ($\bar{X}$), Streuungen (s), Mindest- (Min.) und Maximalzahl (Max.) richtiger zufallskorrigierter Rohwerte der durchgeführten Tests, getrennt nach Schularten

Test	Schulart	$\bar{X}$	s	Zahl d. richtigen Lösungen	
				Min.	Max.
=====					
Gesamtwert	SS	4,42	2,28	1	9
KERN-Schul-	HS	2,24	1,85	0	8
reifetest	RS	1,36	1,34	0	6
(1960/61)	OS	1,18	1,31	0	6
(*)	224	1,87	1,80	0	9
=====					
Gesamtwert	SS	19,67	9,35	10	37
MÜNCHNER-	HS	33,68	10,69	5	59
Schulreife-	RS	40,42	11,20	7	55
test	OS	45,29	10,33	14	64
(1960/61)	224	37,12	12,42	5	64
=====					
Bilderwort-	SS	10,50	4,91	4	21
schatz	HS	16,72	6,17	2	33
(1964)	RS	20,90	6,76	8	36
	OS	25,23	6,22	9	36
	224	19,67	7,49	2	36
=====					
Wortschatz	SS	5,42	5,20	0	17
(1964)	HS	17,70	7,37	0	35
	RS	26,04	5,68	13	37
	OS	29,90	5,69	18	39
	224	22,28	9,24	0	39
=====					
Raumerfassung	SS	6,00	3,36	0	10
(1964)	HS	11,72	4,44	0	24
	RS	13,24	4,01	6	22
	OS	15,71	4,17	9	23
	224	12,86	4,80	0	24
=====					

(*) Bei diesem Test entspricht ein geringer Punktwert "hoher" Schulreife und vice versa.

Fortsetzung der Tabelle 8

Test	Schulart	$\bar{X}$	s	Min.	Max.
=====					
Figurengruppen (1964)	SS	9,33	5,45	0	18
	HS	14,04	4,16	3	23
	RS	16,38	4,30	7	24
	OS	17,42	4,00	6	26
	224	15,25	4,66	0	26

Wortgruppen (1964)	SS	4,09	5,00	0	17
	HS	19,69	7,51	0	36
	RS	24,16	6,15	6	36
	OS	27,03	6,32	8	38
	224	21,88	8,56	0	38

Wahrnehmungsschnelligkeit (1964)	SS	14,17	4,32	8	21
	HS	24,15	7,07	8	46
	RS	26,80	6,13	11	39
	OS	28,60	6,44	13	43
	224	25,44	7,31	8	46

Zahlen (1964)	SS	8,58	4,60	4	13
	HS	26,53	8,74	6	46
	RS	31,76	7,26	13	34
	OS	34,08	7,04	16	35
	224	28,83	9,70	4	46

Wortschatz (1968)	SS	2,08	2,84	0	9
	HS	15,34	7,24	0	35
	RS	24,74	6,69	12	38
	OS	30,11	7,17	14	45
	224	20,82	10,36	0	45

Figurenvergleich (1968)	SS	27,17	10,23	12	47
	HS	36,10	10,78	7	53
	RS	35,00	10,81	13	53
	OS	39,40	12,08	6	53
	224	36,29	11,40	6	53

Fortsetzung der Tabelle 8

Test	Schulart	$\bar{X}$	s	Min.	Max.
=====					
Buchstaben- reihen (1968)	SS	1,50	2,78	0	10
	HS	9,94	5,93	0	23
	RS	14,58	5,14	0	25
	OS	18,87	4,67	7	28
	224	13,44	7,17	0	28
Wortgruppen (1968)	SS	5,83	4,35	0	10
	HS	13,30	5,88	0	31
	RS	17,22	4,70	6	27
	OS	21,21	4,91	11	33
	224	15,96	6,67	0	33
Wahrnehmungs- schnelligkeit (1968)	SS	20,25	5,51	11	27
	HS	34,84	6,36	15	50
	RS	37,52	6,19	25	50
	OS	38,68	6,69	20	50
	224	35,72	7,52	11	50
Zahlen (1968)	SS	4,58	2,07	1	8
	HS	16,91	6,75	0	39
	RS	22,82	9,10	8	43
	OS	22,15	6,75	10	43
	224	19,02	8,38	0	43

Tabelle 9a : Rohwertverteilung im Test "Figurenvergleich",
getrennt für SS, HS, RS und OS.
Gesamtzahl der Items: 53

Rohwert- klassen	SS f	HS f	RS f	OS f	Σf	f %
0 - 3	0	0	0	0	0	0,0
4 - 7	0	1	0	1	2	0,5
8 - 11	0	0	0	0	0	0,0
12 - 15	2	2	1	0	5	2,2
16 - 19	0	2	2	3	7	3,1
20 - 23	2	9	5	4	20	8,9
24 - 27	3	11	6	5	25	11,2
28 - 31	2	10	7	3	22	9,3
32 - 35	0	15	7	8	30	13,4
36 - 39	1	13	4	4	22	9,8
40 - 43	1	6	6	5	18	8,0
44 - 47	1	10	5	5	21	9,4
48 - 51	0	16	2	15	33	14,7
52 - 54	0	5	5	9	19	8,5
N / %	12	100	50	62	224	100

Tabelle 9b : Rohwertverteilung im Test "Buchstabenreihen",
getrennt für SS, HS RS und OS.
Gesamtzahl der Items: 30

Rohwert- klassen	SS f	HS f	RS f	OS f	Σf	f %
0 - 3	11	18	1	0	30	13,4
4 - 7	0	17	1	1	19	8,5
8 - 11	1	22	5	4	32	14,3
12 - 15	0	24	11	9	44	19,6
16 - 19	0	14	19	16	49	21,8
20 - 23	0	5	8	23	36	16,1
24 - 27	0	0	5	8	13	5,8
28 - 30	0	0	0	1	1	0,5
N / %	12	100	50	62	224	100

Tabelle 9c : Rohwertverteilung im Test "Wortgruppen" (1964),
getrennt für SS, HS, RS und OS.
Gesamtzahl der Items: 40

Rohwert- klassen	SS f	HS f	RS f	OS f	Σf	f %
0 - 3	7	1	0	0	8	3,6
4 - 7	3	6	1	0	10	4,5
8 - 11	1	8	1	3	13	5,8
12 - 15	0	15	3	1	19	8,5
16 - 19	1	16	4	1	22	9,8
20 - 23	0	19	10	10	39	17,4
24 - 27	0	21	16	14	51	22,8
28 - 31	0	9	12	14	35	15,6
32 - 35	0	4	2	16	22	9,8
36 - 39	0	1	1	3	5	2,2
N / %	12	100	50	62	224	100

Figurenvergleich in erster Linie durch eine Gruppe von HS verursacht wurde. Eine Tendenz zur Rechtsschiefe geht von einer weiteren Gruppe der HS und der größten Einzelgruppe der OS aus. Beim Test Buchstabenreihen wird die Mehrgipflichkeit offensichtlich durch die SS und zum Teil auch durch HS verursacht. Für diese Schüler scheint der Test zu schwierig zu sein. Darüber hinaus werfen die Ergebnisse die generelle Frage auf, ob es ein Kriterium für einen guten Test sei, wenn die Aufgabenschwierigkeit so steil ansteigt, daß große Gruppen der SS und HS und sogar noch 1 RS nur maximal drei Aufgaben zu lösen vermögen. Die nach Schulart getrennten Rohwertverteilungen des Tests Wortgruppen (1964) in Tabelle 9c weisen - bezogen auf die Gesamtstichprobe - keine Mehrgipflichkeit auf. Die Rechtsschiefe besteht jedoch weiterhin. Auch hier gilt, daß eine Inspektion der Antwortbögen in Verbindung mit Gesprächen mit den Probanden notwendig wäre, um klärende Informationen zu erhalten.

Diese wenigen Darlegungen sollten folgendes deutlich machen:

- 1) Bei der Konstruktion von Fähigkeitstests (aber nicht nur diesen) wird implizit oder explizit von der Normalverteilung der interessierenden

Fähigkeiten ausgegangen bzw. Items und/oder Bearbeitungszeit werden so gewählt, daß sich für eine repräsentative Stichprobe eine Normalverteilung der individuell erzielten Testergebnisse ergibt. Ob z.B. Intelligenz und die in diesem Konstrukt zusammengefaßten Einzelfähigkeiten tatsächlich der Normalverteilungshypothese folgen, bleibt dabei unbeantwortet.

U.W. ist noch nicht abzusehen, ob mit Hilfe der probabilistischen Testtheorie dieser Mangel überwunden werden kann.

- 2) Eine tatsächlich repräsentative Stichprobe stellt eine unverzichtbare Bedingung für die Eichung (Erstellung der Normtabellen) eines Tests dar. Wie die dargestellte Analyse der Verteilungen gezeigt hat, hat sich die große Sorgfalt bei der Erstellung der hier verwendeten beiden Testbatterien PMA 9-10 und PMA 13-14 ausgezahlt. Dennoch konnten (zumindest) bei einigen Tests die (selbst) gestellten Anforderungen (vgl. LIENERT 1967; MEILI & STEINGRÜBER 1978) nicht erfüllt werden, was uns zu dem folgenden Exkurs veranlaßt hat.

7.2

Exkurs: bezogen auf die Eichung einiger bekannter
Intelligenztests, HAWIE, HAWIK, IST, LPS und PSB.

Im Zuge der Bildungswerbung und der Forderung nach Objektivierung der Leistungsmessung sind eine solche Anzahl neuer Tests erschienen, daß allein die große Zahl ein gewisses Unbehagen hinsichtlich der Güte der Tests erzeugt. HOPF (1973) kommt für 1970 in einer Arbeit für die Stiftung Volkswagenwerk zu der Auffassung, "die Herstellung eines guten Schulleistungstests für eine Klassenstufe in einem Unterrichtsfach" koste insgesamt ca. DM 300.000! MICHEL und TROST (1980) geben an, daß für die Entwicklung der "Hochschuleingangstests für Mediziner" bisher "etwa vier Millionen DM" aufgewendet wurden. Darf man davon ausgehen, daß auch nur der größeren Zahl der Testautoren die jeweils notwendigen Mittel zur Verfügung gestanden haben? Und wenn dies so gewesen sein sollte, wie ist es dann zu erklären, daß notwendige Neustandardisierungen vorhandener Tests so selten vorgenommen werden?

Es hat sich mittlerweile durchgesetzt, einem Test eine Handanweisung beizugeben, wie dies z.B. von LIENERT (~1967) gefordert wird. Die Durchsicht der Handanweisungen verschiedener bekannter Tests zeigt, daß die Forderung nicht an Aktualität verloren hat. Sie bedarf geradezu der Unterstützung, denn u.E. sollten Handanweisungen mehr enthalten, als Anwendungshinweise und Normen. Es erscheint uns nützlich, den Testanwender aber auch Testkonstrukteure in der Handanweisung mit allen Problemen vertraut zu machen, mit denen der Testautor konfrontiert wurde, den gelösten wie den (noch) nicht gelösten. Es ist hier nur möglich anzudeuten, welche Fülle von Problemen manche Testautoren den Anwendern aufbürden. Daher kann nur empfohlen werden, Handanweisungen verschiedener Tests nebeneinander zu lesen.

Das Gemeinte soll am Problem der Standardisierung und Eichung der folgenden Intelligenztests verdeutlicht werden (vgl. dazu ANASTASI 1976, S. 67ff):

1. Beispiel: HAWIE (³1964)

In seinem Vorwort zur ersten deutschen Ausgabe des Hawie (1955) schreibt BONDY:

"Der Hamburg-Wechsler-Intelligenztest für Erwachsene wurde zur Standardisierung mehr als 3000 Personen in Westdeutschland gegeben. Die endgültige Stichprobe wurde auf etwa 2000 Personen beschränkt, um eine repräsentative Gruppe zu gewinnen. Wir berücksichtigten dabei die Alters-, Geschlechts- und geographische Verteilung

ten. Als Durchschnitt galt eine Schule, wenn die Schüler als Gruppe weder wesentlich über, noch unter dem vermutlichen Intelligenzdurchschnitt lagen. In der Liste der Schulbehörde waren die verschiedenen Schultypen in annähernd angemessener Anzahl enthalten. ... Im allgemeinen prüften wir an einer Schule nie mehr als 100 bis 150 Kinder, die durch Zufallsverfahren bestimmt wurden" (S. 123 der Handanweisung).

Bezüglich der regionalen Verteilung heißt es auf Seite 119ff:

"Unsere Prüflinge wurden hauptsächlich aus folgenden Gebieten und Orten der Bundesrepublik ausgewählt: Hamburg, Schleswig-Holstein, Bremerhaven, Göttingen und Landgebiete Niedersachsens, Dortmund, Marburg, Frankfurt, Tübingen, München und West-Berlin. Der Hauptanteil der kleinstädtischen und dörflichen Bevölkerung rekrutierte sich aus den Landgebieten Schleswigs-Holsteins und Niedersachsens."

Zu Beginn des Kapitels "Standardisierung und Ergebnisse" heißt es (S. 125ff):

"Da wir nun ca. 3000 Personen untersucht haben, können wir besser eine kompetente Ansicht über die Vorzüge unserer Untertests äußern."

Dazu die Fußnote:

"Diese Zahl enthält etwa 1000 Fälle, die für die Standardisierung nicht benutzt wurden."

Der Testanwender muß den Testautoren beinahe blind vertrauen. Er hat kaum die Möglichkeit, seine eigenen Auffassungen zur Repräsentativität von Stichproben mit denen der Testautoren zu konfrontieren.

Es gäbe noch manchen Aspekt zu diskutieren, wir müssen uns aber an dieser Stelle Beschränkungen auferlegen. Jedoch seien die Normen selbst noch kurz angesprochen. Auf den Seiten 221 bis 232 der Handanweisung finden sich die Tabellen zur Umrechnung der Wertpunkte in IQ-Werte, getrennt für die Altersklassen, wobei in den Altersklassen 10-14 für jede Altersklasse je drei Viermonatsintervalle gebildet wurden. Für die Gruppe der 55 bis 59jährigen beginnen die Wertpunkte zur Ermittlung der Gesamt-IQ's bei 8, fortlaufend bis 155, das sind 148 verschiedene Wertpunkte. In der Eichstichprobe waren aber insgesamt nur 26 Probanden enthalten! Ähnliches gilt auch für die meisten übrigen Altersklassen. Demzufolge können die Normtabellen, jede für sich, kaum das Ergebnis der Transformation empirischer Daten darstellen, sondern sind eher das Produkt primär theoretischer Überlegungen (vgl. dazu den Abschnitt über "Gewichtsbestimmung der einzelnen Untertests" auf Seite 126 der Handanweisung).

So kann es auch nicht verwundern, daß nicht eine einzige Altersgruppe erwartungswidrige Punktwerte produziert.

Beispiel: Da zunächst ein kontinuierlicher Intelligenzanstieg angenommen wird, erhalten Kinder mit gleichem Gesamtpunktwert dafür um so bessere IQ's je jünger sie sind. Ob sich dies aus den empirischen Daten immer so ergeben hat? WECHSLER jedenfalls schreibt in seinem Vorwort zur deutschen Ausgabe des HAWIE (S. 8):

"Die Intelligenz wird im vorliegenden Test durch einen Intelligenzquotienten (IQ) ausgedrückt. Dieser Quotient wird jedoch nicht durch das Verhältnis von Intelligenzalter zu Lebensalter definiert, sondern durch die individuelle Abweichung vom Mittelwert der jeweiligen Altersgruppe."

2. Beispiel: HAWIK (³1966)

Im Vorwort zur 1. Auflage schreibt BONDY zur Stichprobenauswahl:

"Die Standardisierungsstichprobe besteht aus 1500 sechs- bis fünfzehnjährigen Kindern, die für diese Bevölkerungsgruppe in Westdeutschland als repräsentativ angesehen werden."

Und im Vorwort zur 2. Auflage heißt es:

"... haben die Bearbeiter des HAWIK diese neue erweiterte Auflage vorbereitet. Soweit es möglich war, wurden kritische Hinweise von Fachkollegen hierbei berücksichtigt."

Sind solche kritischen Hinweise von Fachkollegen nicht vielfach auch für Anwender interessant, so daß sie auch in einer Handanweisung aufgeführt werden sollten?

Zur Standardisierung des Tests heißt es auf Seite 9:

"Die Standardisierungsstichprobe eines Tests soll so weit wie möglich repräsentativ für die Bevölkerungsgruppe sein, für die der Test vorgesehen ist. Die Normen des HAWIK basieren daher auf Gruppen, die als repräsentativ für 6-15jährige Kinder in der Bundesrepublik angesehen wurden. Da eine unbeschränkte Zufallsauswahl nicht möglich war, wurde die Auswahl durch eine stratifizierte Stichprobe vorgenommen, wobei als Grundlage die Daten des Statistischen Bundesamtes dienten."

Standardisiert wurde die HAWIK pro Altersjahrgang über je 75 Jungen und 75 Mädchen, die ihrerseits hinsichtlich des Wohnorts und der besuchten Schulart repräsentativ für die Bundesrepublik seien. Da jeder Altersjahrgang nochmals in drei Viermonatseinheiten aufgeteilt wurde, standen de facto somit bei Gleichverteilung je Standardisierungseinheit lediglich 50 Probanden zur Verfügung. Zur Erstellung der Wertpunkte heißt es auf Seite 16 der Handanweisung:

"Diese Wertpunkte wurden gewonnen, indem zunächst für jede Altersstufe die Verteilung der Rohpunkte jedes Untertests ermittelt wurde und die üblichen statistischen Größen für sie berechnet wurde. Nach einer Glättung wurden die Rohpunkte jedes Untertests und jeder Altersgruppe in eine Punktskala mit 20 Einheiten transformiert, ..."

In welcher Weise diese "Glättung" notwendig war und wie dabei methodisch vorgegangen wurde (vgl. HELLER & ROSEMAN 1974, S. 103ff), bleibt Geheimnis der Autoren.

Über eine ganze Reihe bisher noch unbearbeitet gebliebener Eichungsfehler, die sich aus der Praxis von Erziehungsberatungsstellen ergeben haben, hat KEMMLER berichtet (KEMMLER 1965, S. 49ff), für Testanweisungsprobleme vgl. KALLENBACH (1970), für Testgebrauch vgl. STOLL (1980).

3. Beispiel: IST (²1955) und IST 70 (⁴1973)
(zitiert wird, wenn nichts anderes angegeben, die 4. Auflage)

Im Vorwort zur 1. Auflage heißt es:

"Mit dem Test, für den Normen vom 14. bis 60. Lebensjahr gegeben werden, wurden bisher über 4000 Personen untersucht."

Und im Vorwort zur 2. Auflage:

"Die Basis für die Normen bilden die Untersuchungsergebnisse von 8642 Personen."

Schließlich im Vorwort zur 3. Auflage:

Es wurden "inzwischen mehr als 1,5 Millionen Personen mit dem I-S-T untersucht ..."

"In die Normierung gehen mehr als 30.000 Ergebnisse ein, und zwar etwa 13.000 für die Berechnung der Schul-Standardwerte."

"In die Tabelle für die verschiedenen Berufsgruppen gehen mehr als 5.000 Fälle ein. Die Normentabellen für die verschiedenen Altersgruppen umfassen mehr als 15.000 Ergebnisse."

Man bekommt den Eindruck, als spreche allein die "große" Zahl der Probanden für die Güte des Tests, wobei jedoch z.B. die Zahl 30.000 nur bei Mehrfachzählungen (!) nachvollzogen werden kann.

Zu den Normen heißt es auf Seite 28 der Handanweisung:

"Eine Repräsentativität der Stichproben wurde angestrebt, aber nicht immer voll erreicht. Für die Normierung der 'Schul-Standardwerte' wurden die Ergebnisse von 12.964 Personen verwendet. Die Berufsgruppenprofile setzen sich aus den Ergebnissen von 5.284 Personen zusammen."

Diese Informationen werden vom Testautor offensichtlich als ausreichend für den Anwender angesehen.

Beim Vergleich der Normen der 2. und 3. Auflage fällt auf, daß - obwohl zur 2. Auflage nur 8642 Probanden zur Verfügung standen - die Standardwerttabellen der 2. Auflage für alle Altersgruppen und alle Untertests SW über die ganzen Rohwertbreiten enthalten. Dieser Sachverhalt muß u.E. doch so interpretiert werden, daß auch die extremen Rohwertleistungen in allen Altersgruppen und für alle Untertests tatsächlich erzielt wurden, zumal ja auch Mittelwerte und Streuungen der Rohwerte jeweils angegeben werden. Obgleich die Stichproben zur Erstellung der Normen für die dritte Auflage beträchtlich vergrößert wurden, stellt man erstaunt fest, daß nunmehr besonders bei den jüngeren Altersgruppen die Höchstleistungen in den Einzeltests wie im Gesamtest nicht mehr erzielt werden. Warum dies so ist, bleibt Geheimnis des Autors.

Der Untertest Gemeinsamkeiten (GE) wurde in der 3. Auflage verändert und die maximal erreichbare Punktzahl von 32 auf 20 reduziert. Dazu erfährt der Anwender im Vorwort zur 3. Auflage, S. 5, lediglich:

"Um eine noch bessere Objektivität zu erreichen, wurde die Aufgabengruppe GE (Gemeinsamkeiten) umkonstruiert."

In der Aufstellung "Vergleich der Geschlechter", S. 33, liest man hierzu:

"Für die neue Version von GE (Gemeinsamkeiten) wurde bei N = 200 ein analoger Vergleich angestellt, der keinen signifikanten Unterschied erbrachte."

Die unter GE eingetragenen Werte sind jedoch noch die der 2. Auflage. Dies gilt auch für alle auf S. 28 unter a) aufgeführten Zuverlässigkeitskoeffizienten (!).

Ebenso willkommen wären dem Anwender sicher Informationen darüber, wie die systematischen Abweichungen der Gesamt-Rohwerte zustandegekommen sind, wie sie Tabelle 10 zeigt.

Schwierigkeiten bestehen auch hinsichtlich der Reliabilitätsangabe. Unter anderm werden in der Handanweisung, Seite 28, Retestkoeffizienten mitgeteilt, von denen der Benutzer jedoch lediglich erfährt, daß den Berechnungen eine Stichprobe von 128 Pbn zugrunde lag, die wechselseitig die Testformen A und B im Abstand eines Jahres bearbeitet hat; keine Angaben über Alter, Geschlecht, besuchte Schule, evtl. Beruf, soziale Schicht etc.. Erst wenn es dem Benutzer gelingt, die dazu vorliegende Publikation einzusehen, erfährt er folgendes:

Bei den Pbn handelt es sich um "Mittelschüler", "Lehrlinge aus einem Werk der chemischen Großindustrie", die bei der ersten Testdurchführung 17 Jahre alt waren (AMTHAUER 1957).

Somit sind über die Reliabilität jedoch nur für einen sehr engen Bereich des Anwendungsspektrums Aussagen möglich. Darauf weisen auch die in der-selben Publikation in gleicher Ausführlichkeit dargestellten Reliabilitätskoeffizienten einer Gruppe 14/15jähriger Lehrlinge hin. Die Koeffizienten liegen alle unter denen der 17/18jährigen; die Gesamtreliabilität beträgt .83 bei den 14/15jährigen und .91 bei den 17/18jährigen, eine nicht unerhebliche Differenz.

Es verwundert nicht, daß AMTHAUER selbst schreibt, die Kenntnis seiner Ergebnisse seien von Bedeutung,

"wenn man einen Test auf seine Brauchbarkeit hin beurteilen will", wozu "die Gruppe der Untersuchten (...) genau charakterisiert werden muß", da sie "in ihrer Zusammensetzung einen entscheidenden Einfluß auf die Höhe der Zuverlässigkeitskoeffizienten hat" (AMTHAUER 1957, S. 165f).

Verwunderung löst allerdings angesichts dieser richtigen Erkenntnis das Fehlen dieser Informationen in der Handanweisung aus. Für den ISI-70 sind u.E. neue - für alle Normgruppen getrennte - Reliabilitätskoeffizienten unbedingt notwendig, da z.B. der Test GE geändert wurde und die Streubreiten der Standardwerte sich für einzelne Altersgruppen nicht unbeträchtlich verschoben haben (vgl. auch ANASTASI 1976, S. 127).

Unser Augenmerk richtete sich hier und für die übrigen Beispiele primär auf "Ungereimtheiten" in den Testmanuals. Auf andere Probleme der Testkonstruktion haben jüngst SIMONS und LANGHEINRICH in ihrer Fehleranalyse

Tabelle 10 : Mittelwerte (M) und Differenzen der Gesamt-
rohwertverteilungen des IST und IST 70 für die
verschiedenen Altersgruppen und Schulabschlüsse
der Eichstichproben

IST u. IST 70	12	13	14	15	16	17	18
M _{IST 70}	59	65	70	77	83	88	91
M _{IST}	—	63	66	73	81	85	89
Differenz	—	+2	+4	+4	+2	+3	+2

t-Test für abhängige Stichproben: $p < 0,1\%$;
N = 6

IST 70	19-20	21-25	26-30	31-35	36-40	41-45	46-50	51-60
IST	19-20	21-24	25-28	29-33	34-39	40-45	46-60	46-60
M _{IST 70}	93	92	90	87	83	78	73	66
M _{IST}	98	97	96	95	90	85	79	79
Differenz	-5	-5	-6	-8	-7	-7	-6	-13

t-Test für abhängige Stichproben: $p < 0,1\%$;
N = 7

IST u. IST 70	Hauptschul- abschluß	Mittlere Reife	Abitur
M _{IST 70}	77	102	116
M _{IST}	82,2	107,4	120
Differenz	-5,2	-5,4	-4

t-Test für abhängige Stichproben: $p < 1\%$;
N = 3

zum Analogietest des IST-70 beispielhaft aufmerksam gemacht (SIMONS & LANGHEINRICH 1980). SADER (1961, S. 74) weist auf die Begrenzungen der Testinterpretation hin. Schließlich sei an dieser Stelle noch angemerkt, daß auch Testkompendien teilweise eher unkritisch Angaben der Testautoren wiedergeben und auf mögliche Ergänzungen verzichten (HILTMANN 1977; SCHMIDT-CHEN 1975; WEISE 1975). Dadurch wird leider die Funktion eines Korrektivs, die Kompendien neben der reinen Nachschlagefunktion haben sollten, verschenkt (vgl. BÜROS 1975).

4. Beispiel: LPS (1962) und PSB (1969)

Für das LPS liegen Normen für 14 Altersklassen zwischen 10 - 50 Jahren und älter, und für das PSB Normen für 23 Altersklassen zwischen 9 - 20 Jahren vor. Das PSB besteht aus den Teilstests des LPS, welche

"für schulische Zwecke besonders wichtig sind. ... Diese 10 Subtests werden im vorliegenden PSB auf einem DIN-A 4-Bogen untergebracht. Prüflinge und Prüfer sparen dadurch viel Zeit. Dennoch sind die Ergebnisse von großer diagnostischer Relevanz. Die besonders sorgfältige umfassende Eichung, ..." (Vorwort zum PSB)

Zur Eichstichprobe heißt es im LPS, S. 6:

"Die letzte Eichung stützt sich auf eine an Hand von über 5300 Prüfergebnissen ermittelte Einteilung der elterlichen Berufe in 5 Klassen ... Allein mit den Bogen der letzten Auflage wurden zu Eichzwecken vom Verfasser über 3000 Vpn getestet. Die volle Arbeitskurve wurde jedoch im allgemeinen nur bei Berufsschülern und Kriminellen ermittelt."

Fragen zur Repräsentativität werden nicht erwogen. Bei Annahme der Gleichverteilung standen $5300 : 14 = \text{ca. } 379$ Pbn je Altersgruppe zur Verfügung. Diese Zahl bezieht sich jeweils auf 2 Paralleltests. Wie die Datensammlung tatsächlich vorgenommen wurde, dazu gibt es im Vorwort zum LPS noch einen Hinweis:

"In den Jahren 1961/62 erfolgte schließlich die Eichung. Zu dieser schickten auch die Psychologischen Institute der Universitäten Freiburg, Hamburg, Mainz und Saarbrücken LPS-Ergebnisse in größerer Zahl."

Zur Eichung des PSB heißt es auf Seite 17 der Handanweisung:

"Die Landesanstalt für Erziehung und Unterricht ... in Stuttgart ließ in Dörfern und Städten von Baden-Württemberg nicht nur umfangreiche Untersuchungen mit

dem LPS durchführen, sondern erhob auch an rund 10.000 Schülern die Ergebnisse für die Eichstichprobe des PS8. Für die Altersgruppen 10-10;3, 10;4-10;7, ... usw. wurden vom Testautor jeweils aus dem größeren Kollektiv repräsentative Stichproben von je N=200 zusammengestellt (bei 17jährigen N=150, ...). Jede Stichprobe enthielt je 50 Jungen und 50 Mädchen bei Form A und B, d.h. also je 100 Jungen und Mädchen."

Dabei wurde entsprechend den Gegebenheiten in Baden-Württemberg für jede "repräsentative Stichprobe" sichergestellt, daß Probanden nach

a) regionalen Aspekten (Großstadt, Mittelstadt, Kleinstadt, Dorf);

b) Beruf der Väter (5 Gruppen);

c) Schulart ab 12 Jahren (Haupt-, Mittel- und Oberschüler, jedoch keine Sonderschüler);

enthalten waren. Das hier gewählte Quotenverfahren, nämlich nur dafür zu sorgen, daß jede Einzelstichprobe insgesamt die relativen Merkmalanteile aufweist, schließt allerdings nicht aus, daß Verzerrungen vorkommen; z.B. Jungen, die zur Oberschule gehen, deren Väter Hochschulabschluß haben und die in einer Großstadt wohnen, könnten über- oder unterrepräsentiert sein. Dies wiederum kann zu fragwürdigen Normtabellen führen.

Um eine Vorstellung davon zu erhalten, wie tatsächlich für jede Altersgruppe - entsprechend den vom Testautor als relevant erachteten Merkmalen - repräsentative Eichstichprobe zusammengesetzt und welchen Mindestumfang sie haben müßten, kann folgende Rechnung hilfreich sein:

Wenn auch nur je 1 Proband mit den relevanten Merkmalskombinationen in jeder Eichstichprobe enthalten sein soll, so ergibt sich folgendes:

Relevant sind lt. Testautor:

- 2 Geschlechter,
- 5 Berufsgruppen,
- 4 Ortsgrößen,
- 3 Schularten,
- 22 Altersgruppen,
- 2 Paralleltestformen.

Diese Zahlen miteinander multipliziert ergibt 5280 Probanden, d.h. bei 10.000 Probanden in der Eichstichprobe kann jede notwendige Merkmalskombination insgesamt nicht einmal mit zwei Pbn vertreten sein. Um in jeder Altersgruppe und beiden Testformen A und B auch nur jeweils 1 Träger mit den als relevant erachteten Merkmalen zu haben, müßte jede vom Testautor als repräsentativ bezeichnete Stichprobe nicht 200 sondern 240 Probanden enthalten. Das vom Testautor gewählte

Quotenverfahren, das ja nicht unumstritten ist (vgl. SELG und BAUER 1971, S. 92f), bereitet somit auch Schwierigkeiten für die Zusammenstellung der Stichprobe. Für testvergleichende Kritik siehe GREIF (1972).

Wir brechen hier den Exkurs unserer Gegenüberstellungen ab. Es ist unverkennbar, daß die Testeichungen Fragen aufwerfen, auf die dem Testbenutzer in den Handanweisungen Antworten bereit gestellt werden müßten. Es hat uns sehr überrascht, in den Handanweisungen so wenig z.B. über Verteilungsanalysen der Testrohwerte jeder Alters-Normgruppe zu erfahren. Diese Kritik kann u.E. nicht etwa mit einem Verweis auf die Literatur zurückgewiesen werden. Da die Testergebnisse häufig weitreichende Konsequenzen für die Probanden nach sich ziehen, besonders wenn die Ergebnisse negativ sind, erscheint es notwendig, daß die Fachverbände sich dieser Problematik annehmen. Sie sollten festlegen, was alles in Handanweisungen festgehalten werden muß. Und sie sollten ein "Gütesiegel" entwickeln, daß den Status eines Tests, seinen Anwendungsbereich, Änderungserfordernisse, usw. beschreibt (vgl. auch BRICKENKAMP 1981, S. 134). Da die Testentwicklung nicht zuletzt eine Kostenfrage ist, sollten mit Verweis auf die Kosten für die Hochschulzugangstests Politiker und Forschungsförderungsinstitutionen dafür gewonnen werden, für Testentwicklungen - sofern man derartige Instrumente für notwendig hält - ganz gezielt Mittel zur Verfügung zu stellen. Gerichtsentscheidungen in den USA (vgl. z.B. DER SPIEGEL vom 17. 12. 1979) lassen vermuten, daß es lediglich eine Frage der Zeit ist, bis der Einsatz von Tests, die in Konstruktion und Eichung soweit hinter den theoretischen Erfordernissen und praktischen Möglichkeiten zurückbleiben, auch hierzulande eingeschränkt wird. Schwerwiegende Angriffsflächen dazu bieten u.E. eine Reihe unserer Tests ohne Zweifel.

Um jedoch keine falschen Schlußfolgerungen zu provozieren, sei gleich angemerkt, daß die angeführten Mängel kein Spezifikum deutscher Tests darstellen. Vielmehr finden sich gleiche und/oder ähnliche Probleme auch bei sehr bekannten amerikanischen Tests wie z.B. den Primary Mental Abilities (THURSTONE 1965), dem Wechsler Intelligence Scale for Children - WIS-R - (WECHSLER 1974), dem Wechsler Adult Intelligence Scale - WAIS - (WECHSLER 1965) oder dem Otis-Lennon Mental Ability Test (Otis & Lennon 1969). Dieses Defizit ist gerade bei amerikanischen Tests nur schwer zu verstehen, hat doch der amerikanische Psychologenverband (American Psychological Association, APA) bereits 1954 entsprechende Standards verabschiedet (APA 1974).

Mißt man die in Testmanuals dargebotenen Informationen am theoretisch Machbaren und praktisch Wünschbaren, so bleibt eine beträchtliche Anzahl nicht unbilliger Fragen offen. Auch für Tests sollte aber einzulösen versucht werden, was METZGER für empirische Untersuchungen gefordert hat.

7.3 Objektivität, Reliabilität, Validität

Zunächst sei hierzu auf die Ausführungen bei KEMMLER (1967, S. 37ff; 1976, S. 39ff) verwiesen.

7.4 Objektivität

Objektivität ist ein Maß für den Grad der intersubjektiven Übereinstimmung hinsichtlich Testdurchführung, Testauswertung und Testinterpretation (vgl. z.B. HELLER/NICKEL/ROSEMAN 1978, S. 89ff). Im Kapitel zur formalen Auswertung der Tests und Fragebögen (s. S.) wurden bereits Angaben gemacht, die die Objektivität der Tests betreffen. Die beiden Schulreifetests lassen in dieser Hinsicht einige Wünsche offen, zumindest bezüglich der Auswertung. Bei den PMA 9-10 und PMA 13-14 treten keine Probleme auf, da die meisten Tests nur Aufgaben nach dem Multiple-choice-Verfahren enthalten. Die Aufgaben der Rechentests haben alle eindeutige Lösungen, so daß auch hier die Bedingungen für eine objektive Testauswertung erfüllt sind. Bezüglich der Testdurchführung kann gesagt werden, daß Versuchsleiter und -helfer in Vorversuchen auf ihre Aufgaben im Sinne der Durchführungsanleitungen geschult wurden. Ob nicht dennoch Versuchsleitereffekte aufgetreten sind, kann anhand unserer Daten kaum ermittelt werden, da dies eine zufällige Ver-

teilung der Untersuchungsgruppen erfordert hätte, was jedoch technisch häufig nicht möglich war.

7.5 Reliabilität

Die Reliabilität liefert ein Maß für den Grad der Genauigkeit, mit der ein Test das mißt, was er faktisch mißt. Reliabilitätskoeffizienten, die von KEMMLER (1967, 1976) ermittelt wurden, basieren auf der Testhalbierungsmethode. Über damit im Zusammenhang stehende Probleme sei auf MAGNUSSON (1969, S. 116-127), für neuere Ansätze auf RAATZ (1968, S. 159-175) verwiesen.

Die folgende Tabelle 11 enthält die von KEMMLER berechneten Halbierungskoeffizienten, aufgewertet nach SPEARMAN-BROWN, und die von anderen Testautoren übernommenen. Dazu noch eine Anmerkung zum Begriff der Stabilität von Meßwerten. KEMMLER (1976, S. 39f) verwendet den Begriff - wie z.B. DORSCH (⁹1976, S. 581) - im Zusammenhang mit Halbierungskoeffizienten. Da die Halbierungsreliabilität nur auf einmaliger Testdurchführung und Aufteilung der Testaufgaben in zwei gleichwertige Testhälften (die Aufgaben mit ungerader lfd. Nummer werden häufig als erste Testhälfte und diejenigen mit gerader lfd. Nummer als zweite Testhälfte bezeichnet) basiert, drückt ein so erhaltener Koeffizient die innere Konsistenz beider Testhälften aus. Der Begriff

Tabelle 11 : Reliabilitätskoeffizienten der
2 Schulreifetests, der 7 PMA 9-10
und der 6 PMA 13-14

Test	N	r_{tt}	Methode bzw. Quelle
=====			
Gesamtwert KERN-Schulreifetest	?	?	Keine Angaben bei KERN (1957)
=====			
Gesamtwert MÜNCHNER- Schulreifetest		0,94	Keine Angaben bei WALTER (o.J.)
=====			
Bilderwortschatz	251	0,87	Halbierungsmethode ⁺)
=====			
Wortschatz	251	0,92	"
=====			
Raumerfassung	251	0,72	"
=====			
Figurengruppen	251	0,64	"
=====			
Wortgruppen	251	0,89	"
=====			
Wahrnehmungs- schnelligkeit		0,98	THURSTONE 1975
=====			
Zahlen		?	HYLLA & KRAAK 1954
=====			
Wortschatz	224	0,76	Halbierungsmethode ⁺)
=====			
Figurenvergleich		0,79	THURSTONE & THURSTONE (1958)
=====			
Buchstabenreihen	224	0,91	Halbierungsmethode ⁺)
=====			
Wortgruppen	224	0,72	"
=====			
Wahrnehmungs- schnelligkeit	224	0,98	"
=====			
Zahlen		0,93	DÜKER & LIENERT (1959)
=====			

⁺) vgl. KEMMLER, 1967, S. 37f; 1976, S. 39f

der Stabilität bezieht sich demgegenüber mehr auf Retestmessungen, d.h. auf die (mit zeitlichem Abstand) zweimalige Durchführung desselben Tests an derselben Stichprobe, wobei dann der aus diesen Messungen zu berechnende Korrelationskoeffizient angibt, wie stabil oder konstant das gemessene Merkmal bei sonst gleichen situativen Bedingungen (z.B. Ausschluß von Lerneffekten) ist (vgl. z.B. LIENERT ²1967, S. 218; MAGNUSSON 1969, S. 128; PAYNE und McMORRIS ²1975, S. 384; AMERICAN PSYCHOLOGICAL ASSOCIATION 1974).

Sofern man über Tests verfügt, die den Anforderungen hinsichtlich Itemanalyse (z.B. optimaler Aufgabenschwierigkeit und -trennschärfe) für Niveautests genügen, ist die Berechnung der Halbierungsreliabilität die Methode der Wahl, weil hier die meisten situativen Einflüsse, die bei Paralleltest- und Retestmessungen fast unvermeidbar auftreten, nicht wirksam werden können. Problematisch wird die Berechnung der Halbierungsreliabilität bei Tests, die eine starke Schnelligkeitskomponente haben, wie dies bei den Tests Wahrnehmungsschnelligkeit und Zahlen der Fall ist.

Wie bereits bei der Stichprobenbeschreibung ausgeführt, bestand das Primärinteresse KEMMLER's nicht so sehr darin, Informationen über einzelne Schüler,

sondern über spezifische Schülergruppen zu gewinnen. Nur aus diesem Grunde schien es KEMMLER vertretbar, für die beiden PMA-Batterien auch Gesamt-Reliabilitäten zu ermitteln, die für einzelne Tests auf verschiedenen Berechnungsmodi beruhen, bzw. verschiedenen Quellen entnommen wurden (vgl. hierzu S.).

KEMMLER (1976, S. 40f) hat bereits eine Wertung der Reliabilitätskoeffizienten vorgenommen, der wir uns in den kritischen Akzenten anschließen. Es sei an dieser Stelle nochmals auf unsere detaillierte Beschreibung der Stichprobe zu den verschiedenen Untersuchungsterminen verwiesen. Die teilweise sehr bedeutsamen Abweichungen bei der Merkmalsverteilung einzelner Probandengruppen, die ihrerseits die Streuungen der Testrohwerte (systematisch) mitbestimmen, können die von KEMMLER berechneten Reliabilitätskoeffizienten ebenfalls beeinflusst haben, daß also "die Reliabilität als Varianzverhältnis stark von der Verteilung der Maßzahlpopulation abhängt" (FISCHER 1968, S. 34).

Zur Reliabilität des IST 70 sei auf die Handanweisung (AMTHAUER ⁴1973) verwiesen.

7.5.1 Standardisierung der Tests

Da sich die inhaltlichen Analysen hier immer auf die für die Untersuchung 1968 noch erreichte Stichprobe von $N = 224$ beziehen und wir der Meinung sind, daß die Stichprobenschrumpfung zwar zufällig ist, aber die Normalverteilungsprüfungen bei einigen Tests Abweichungen erkennen lassen haben, haben wir die zufallskorrigierten Rohwerte T -transformiert¹⁾ (nach McCALL) über die beiden Altershalbjahre. Außerdem wurde eine T -Transformation auch über $N = 224$ vorgenommen. (Es wird im Text jeweils angegeben, welche Standardwerte verwendet wurden.) Für Tests, deren Rohwerte normalverteilt sind, ist die T -Transformation mit der z -Transformation identisch. Neben den Verteilungsüberlegungen schien eine Standardisierung auch wegen der unterschiedlichen Itemzahlen notwendig, damit die Werte aus verschiedenen Tests leichter miteinander vergleichbar sind.

Die berechneten Standardwerte (SW, USW, GSW):

- 1) Standardwerte (SW) für die Gesamtrohwerte der beiden Schulreifetests;
- 2) Standardwerte (SW) für die 7 Schulreifetests der PMA 9-10;

¹⁾ Dem Kollegen Jochen WINDHEUSER sei für die Überlassung seines EDV-Programms herzlich gedankt.

- 3) Unterstandardwerte (USW) für
 - den Faktor "Verbal" aus den Standardwerten der Tests Bilderwortschatz und Wortschatz;
 - den Faktor "Reasoning" aus den Standardwerten der Tests Figurengruppen und Wortgruppen;
 - den Faktor "Lesefrei" aus den Standardwerten der Tests Bilderwortschatz, Raumerfassung, Figurengruppen, Wahrnehmungsschnelligkeit und Zahlen;
- 4) Gesamtstandardwert (GSW) über die Standardwerte der 7 PMA 9 - 10;
- 5) Standardwerte (SW) für die 6 Subtests der PMA 13 - 14;
- 6) Unterstandardwerte (USW) für
 - den Faktor "Reasoning" aus den Standardwerten der Tests Buchstabenreihen und Wortgruppen;
 - den Faktor "Lesefrei" aus den Standardwerten der Tests Figurenvergleich, Wahrnehmungsschnelligkeit und Zahlen;
- 7) Gesamtstandardwert (GSW) über die 5 Subtests der PMA 13 - 14.

7.5.2 Standardmessfehler

Wer Informationen über die Genauigkeit einer Messung haben möchte, muß etwas über den Fehler, mit dem jede Messung behaftet ist, wissen. Dazu steht eine einfache Formel bereit, deren wesentlichste "Größe"

die Reliabilität r_{tt} ist. Dazu kommt die Standardabweichung (Streuung) s_x der Messwertverteilung (vgl. MAGNUSSON 1969, S. 86 - 97; ANASTASI ⁴1976, S. 127 - 129).

Die vollständige Formel für den Standardmessfehler (s_e) lautet:

$$s_e = s_x \sqrt{1 - r_{tt}}$$

In der folgenden Tabelle 12 geben wir eine Zusammenstellung der relevant erscheinenden Standardmessfehler wieder.

Tabelle 12 : Standardmessfehler (s_e) der Unter- und Gesamtstandardwerte der Untersuchungen 1964 und 1968

Tests	r_{tt}	s_e (p < 5%)
USW-Verbal 1964	0,94	$\pm$ 4,8
USW-Reasoning 1964	0,89	$\pm$ 6,5
USW-Lesefrei (5T.) 1964	0,92	$\pm$ 5,5
GSW 1964	0,94	$\pm$ 3,9
USW-Verbal 1968	0,76	$\pm$ 9,6
USW-Reasoning 1968	0,91	$\pm$ 5,9
USW-Lesefrei (3T.) 1968	0,98	$\pm$ 2,8
GSW 1968	0,95	$\pm$ 4,4

7.5.3 Interkorrelationen der Testwerte

Bei der Beschreibung der verwendeten Testverfahren wurde dargelegt, daß die PMA 9 - 10 und PMA 13 - 14 von KEMMLER in Analogie zur Faktorentheorie THURSTONE's entwickelt bzw. ausgewählt wurden. Im folgenden sollen nun die Interkorrelationen der PMA und der Schulreife-tests dargestellt werden. Als wir die Berechnungen durchführten, ließen wir uns bei der Wahl der Prüftests vom Stand der Diskussion in einschlägigen Statistikbüchern leiten. Danach erfordert die Berechnung von Produkt-Moment-Korrelationskoeffizienten¹⁾ Normalverteilung der Merkmale, Intervallskalenniveau und bestimmte Mindestgrößen der Stichproben der zu vergleichenden Meßwertreihen. Da vorherige Verteilungsprüfungen häufig zu aufwendig schienen, haben wir uns im Zweifelsfalle für die Berechnung von Rangkorrelationen nach SPEARMAN¹⁾ (mit Korrektur bei Rangplatzteilungen) entschieden. Vergleichsrechnungen über dieselben Variablen und Probanden haben jedoch für Produkt-Moment- als auch für Rang-Korrelationen auch bei ziemlich anomalen Verteilungen keine bedeutsamen Abweichungen ergeben, lediglich die Tendenz, daß die Rangkorrelationen meist etwas niedriger (jeweils in Richtung Null) ausfielen, was ja auch erwartet werden kann.

1) Den Kollegen Dieter LANGHEINRICH und Ulrich ZINGELER sei für die Überlassung ihrer EDV-Programme herzlich gedankt.

Inzwischen liegt eine Arbeit von HAVLICEK & PETERSON (1977) vor. Die Autoren haben gefunden, daß selbst bei sehr starker Verletzung der Grundannahmen das Prüfverfahren von PEARSON angemessen verwendet werden könne.

In der Literatur, besonders im Kontext der Reliabilität, finden sich häufig - mit Abbildungen untermauert - Ausführungen, wonach die Reliabilität eines Tests bzw. die Korrelation zweier Meßwertreihen direkt von der (den) Streuung(en) der Meßwerte abhängt.

So stellt LIENERT (²1967, S. 237) z.B. fest:

"Einen sehr wesentlichen und kaum zu unterschätzenden Einfluß auf die Reliabilitätskennwerte hat auch die Streuung der Testrohwerte. ... Aus der Abb. 24 ist ohne weiteres ersichtlich, daß mit zunehmender Testwertstreuung der Reliabilitätskoeffizient anwächst."

BELSER (³1975, S. 148f) schreibt dazu in seinem Buch "Testentwicklung":

"Die Abhängigkeit der Zuverlässigkeit eines Tests von der Untersuchungsgruppe läßt sich anhand folgender Überlegungen gut demonstrieren. Falls wir in dem von uns angeführten Beispiel ... getrennte Zuverlässigkeitskoeffizienten für Volksschüler und für Gymnasiasten berechnen, erhalten wir unterschiedliche Werte. Beide werden niedriger sein als der Koeffizient für die gesamte Gruppe, weil der Koeffizient abhängig ist von der Streuung der gemessenen Fähigkeit. ... Je homogener die Untersuchungsgruppe (nachgewiesen an der geringeren Streuung, ausgedrückt durch eine geringere Standard-Abweichung), desto niedriger ist der Korrelationskoeffizient für den gleichen Test."

Diese Darstellung halten wir für unzutreffend. Der Korrelationskoeffizient (nach PEARSON) ist definiert als

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

wobei s_{xy} der Kovarianz von X und Y und s_x der Streuung von X bzw. s_y derjenigen von Y entspricht, d.h. der Korrelationskoeffizient ist eine Verhältniszahl und wird demzufolge von der Kovarianz und den Streuungen der Meßwertreihen determiniert. Dabei können sich Kovarianz und Streuungen durchaus unterschiedlich, ja sogar unabhängig voneinander verändern. Wäre es anders, würde die Definition unzweckmäßig.

Ein Beispiel soll dies verdeutlichen:

Vollständige Korrelation (+1 oder -1) ist nur zu erzielen, wenn die Kovarianz (absolut) ihr Maximum erreicht, das gleich dem jeweiligen Produkt der Streuungen der Meßwertreihen ist. Ebenso ist eine Nullkorrelation nur dann gegeben, wenn die Kovarianz Null ist (nicht wenn eine Streuung oder beide Null sind).

Gegeben seien N = 12 Probanden, die - jede für sich - in zwei Tests jeweils vollständig identische Meßwerte von 1 - 12 erzielt haben (Pb 1: X = 1, Y = 1; Pb 2: X = 2, Y = 2; ...; Pb 11: X = 11, Y = 11 und Pb 12: X = 12, Y = 12). In diesem Falle ist die Korrelation maximal, also +1, denn die Kovarianz ist gleich dem Produkt der Streuungen, die für beide Variablen in diesem Falle gleich sind.

Im zweiten Fall sollen sich gleiche Meßwerte ergeben haben, mit dem einen Unterschied, daß der Proband 1 im zweiten Test den Wert 12 und der Proband 12 den Wert 1 erreicht habe, also

Pb 1: $X = 1$, $Y = 12$; Pb 2: $X = 2$, $Y = 2$; Pb 3: $X = 3$, $Y = 3$;
...; Pb 11: $X = 11$, $Y = 11$; Pb 12: $X = 12$, $Y = 1$.

Die Streuungen der Meßwertreihen X und Y sind natürlich völlig identisch mit denen im ersten Fall. Was sich ändert, ist allein die Kovarianz, sie wird erheblich geringer. Die Folge davon ist eine erhebliche Verminderung des Korrelationskoeffizienten. Mit anderen Worten: Korrelationskoeffizienten können - unter sonst gleichen Bedingungen - bei gleichen Streuungen in verschiedenen Meßwertreihen verschieden ausfallen, wenn nur die Covarianz sich ändert.

Dieser Hinweis scheint berechtigt, da u.E. allzu oft sog. Minderungskorrekturen vorgenommen werden, wenn eine mehr oder weniger homogene Stichprobe, bezogen auf demographische Merkmale gebildet wird, ohne das in jedem Falle klar ist, daß die Homogenisierung die Korrelation zwischen psychischen Merkmalen auch tatsächlich mindert (vgl. TENT 1969, S. 32).

a) Interkorrelationen der Testrohwerte

aa) Gesamtstichprobe, N = 224

Tabelle 13 enthält die Interkorrelationen der Testrohwerte von 15 kognitiven Variablen für die Gesamtstichprobe.

Tabelle 13 :

Interkorrelationen (PEARSON) der Rohwerte der beiden
Schulreife tests 1960 (Tests 1+2), der 7 Subtests der PMA 9-10
(Tests 3-9) und der 6 Subtests der PMA 13-14 (Tests 10-15)
Gesamtstichprobe (SS + HS + RS + OS), N = 224

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1														
2	.66													
3	.22	.40												
4	.31	.49	.70											
5	.33	.48	.45	.52										
6	.22	.41	.35	.45	.46									
7	.27	.50	.46	.65	.51	.56								
8	.33	.48	.36	.40	.45	.44	.44							
9	.24	.39	.36	.49	.45	.41	.55	.57						
10	.31	.48	.73	.81	.54	.47	.64	.36	.46					
11	.24	.26	.26	.20	.45	.32	.23	.31	.25	.18				
12	.37	.50	.49	.65	.53	.54	.62	.51	.53	.65	.35			
13	.21	.44	.48	.66	.55	.53	.66	.41	.47	.67	.26	.65		
14	.36	.47	.40	.45	.43	.39	.46	.70	.54	.39	.36	.56	.37	
15	.18	.33	.28	.40	.34	.33	.44	.34	.57	.38	.32	.46	.44	.47

- 1 = Gesamtrohwert-KERN
4 = Wortschatz
7 = Wortgruppen
10 = Wortschatz
13 = Wortgruppen
- 2 = Gesamtrohwert-MÜNCHNER
5 = Raumerfassung
8 = Wahrnehmungsschnelligkeit
11 = Figurenvergleich
14 = Wahrnehmungsschnelligkeit
- 3 = Bilderwortschatz
6 = Figurengruppen
9 = Zahlen
12 = Buchstabenreihen
15 = Zahlen

Signifikanzgrenzen der Korrelationskoeffizienten:

r > .14 p < 5%; r > .18 p < 1%

Die Hauptergebnisse:

Zunächst fällt der durchgehend geringe - wenngleich sehr signifikante - Zusammenhang des Gesamtwerts-KERN mit allen PMA-Untertests auf. Dagegen sind, abgesehen von den Zahlentests und Figurenvergleich, die Beziehungen zwischen dem MÜNCHNER-Schulreifetest und den Untertests der PMA über 4 und 8 Jahre von etwa gleicher Größenordnung (r : 0,40-0,50), aber beträchtlich höher als die analogen Werte im "KERN".

Die Interkorrelationen der PMA 9 - 10 liegen zwischen 0,35-0,70. Insgesamt sind sie aber zu hoch, um von unabhängigen Faktoren im Sinne THURSTONE's sprechen zu können. Ähnliches gilt auch für die PMA 13-14; hier schwanken die Koeffizienten zwischen 0,18-0,67, wobei alle Werte unter 0,37 sich auf Vergleiche mit dem Test Figurenvergleich beziehen, der ja zu den weniger gut konstruierten Tests zählt.

Einigermaßen zufriedenstellend sind die Interkorrelationen analoger Tests der beiden Testbatterien: Die Ergebnisse der 14jährigen im Wortschatz korrelieren mit denen der 10jährigen 0,81 bzw. 0,73 mit Bilderwortschatz. Dies entspricht einer gemeinsamen Varianz bzw. einem Determinationskoeffizienten ($= r^2 \cdot 100$) von 66% bzw. 53%. Die Tests Raumerfassung und Figurenvergleich, die den Faktor "Space"

messen sollen, korrelieren nur mit 0,45 bzw. haben nur 20% gemeinsame Varianz. Die Interkorrelationskoeffizienten der Tests Figurengruppen und Wortgruppen (1964) sowie Buchstabenreihen und Wortgruppen (1968), die den Faktor "Reasoning" messen sollen, schwanken zwischen 0,53-0,66, resp. 28% - 44% gemeinsamer Varianz. Die Korrelation zwischen den beiden Wahrnehmungsschnelligkeitstestergebnissen (Faktor "Perceptual Speed") beträgt 0,70, resp. 49% gemeinsamer Varianz. Schließlich korrelieren die beiden Zahlentests (Faktor "Number") mit 0,57, resp. 32% gemeinsamer Varianz.

Herausragend sind drei Ergebnisse:

- 1) Der Wortschatztest von 1968 korreliert mit .81 auffallend hoch mit dem Wortschatztest von 1964 und damit beträchtlich höher als die beiden Tests Wortschatz und Bilderwortschatz von 1964 untereinander. Die Fähigkeiten im Bereich des Verbalen erweisen sich damit als (zeitlich) relativ stabil.
- 2) In die gleiche Richtung deutet die Korrelation von 0,66 des Tests Wortgruppen (1968) mit dem Test Wortgruppen (1964). Auch hier ist der Zusammenhang dieser beiden Tests höher als derjenige der Tests Wortgruppen und Figurengruppen des Jahres 1964 mit 0,56 bzw. gleich hoch wie derjenige der Tests Wortgruppen und Buchstabenreihen des Jahres 1968 mit 0,65. Damit scheint auch die

Fähigkeit im Bereich des schlußfolgernden Denkens (Reasoning), wie sie besonders mit dem Test Wurtgruppen (1968) gemessen wird, (zeitlich) relativ stabil zu sein.

- 3) Der besonders hohe Stellenwert des Verbalen wird dadurch deutlich, daß der Wortschatztest von 1968 mit allen Tests, die eine verbale Komponente haben, gleich hoch oder sogar höher korreliert als die Tests, die zu den beiden Untersuchungszeitpunkten 1964 und 1968 je spezifische Faktoren (Verbal und Reasoning) messen sollten.

ab) Hauptschüler, N = 100

Tabelle 14 enthält über der Diagonalen die Interkorrelationen der Testrohwerte von 15 kognitiven Variablen für die Gruppe der Hauptschüler.

Die Hauptergebnisse:

Zunächst fällt auf, daß zwischen dem Schulreifetest von KERN und allen Untertests der beiden PMA-Batterien kein bedeutsamer Zusammenhang besteht. Beim MÜNCHNER-Schulreifetest ergeben sich die höchsten Korrelationen von jeweils 0,30 mit den Tests Raumerfassung und Wahrnehmungsschnelligkeit, den Tests von 1964 also, die offenbar Fähigkeiten ansprechen, die beim Erlernen des Lesens und Schreibens notwendig sind (vgl. unten RS + OS).

Interkorrelationen (PEARSON) der Merkmale der beiden
Schulreifetests 1960 (Tests 1+2), der 7 Subtests der PMA 9-10 (Tests 3-9)
und der 6 Subtests der PMA 13-14 (Tests 10-15)

Über der Diagonalen: Hauptschüler, N = 100
unter der Diagonalen: Real- und Oberschüler, N=112

Tests	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	—	.53	-.03	-.06	.08	.02	.01	.08	.08	-.10	.11	.08	-.16	.07	-.11
2		.61	—	.14	.15	.30	.12	.28	.30	.02	.19	.06	.14	.26	.04
3			.22	—	.50	.30	.17	.20	.19	.05	.53	.21	.20	.17	-.11
4				.20	.30	.62	—	.34	.14	.42	.22	.23	.56	.10	.29
5					.30	.37	.29	.29	.31	.41	.33	.31	.38	.45	.46
6						.41	.18	.36	.35	—	.36	.33	.44	.23	.37
7							.33	.29	.42	.24	.52	—	.32	.47	.44
8								.42	.20	.09	.33	.21	.14	.52	.10
9									.24	.15	.07	.19	.03	.08	.31
10										.12	.18	.66	.70	.35	.30
11											.32	.34	.18	.14	.39
12												.25	.40	.24	.39
13													.12	.31	.36
14														.35	.34
15															.01

- 1 = Gesamtrahwert-KERN
4 = Wortschatz
7 = Wortgruppen
10 = Wortschatz
13 = Wortgruppen
2 = Gesamtrahwert-MÜNCHNER
5 = Raumerfassung
8 = Wahrnehmungsschnelligkeit
11 = Figurenvergleich
14 = Wahrnehmungsschnelligkeit
3 = Bilderwortschatz
6 = Figurengruppen
9 = Zahlen
12 = Buchstabenreihen
15 = Zahlen

Signifikanzgrenzen der Korrelationskoeffizienten für HS und RS+OS:

r > .20 p < 5%; r > .26 p < 1%

Die Interkorrelationen der PMA 9 - 10 liegen zwischen 0,05 (Bilderwortschatz und Zahlen) und 0,52 (Wahrnehmungsschnelligkeit und Zahlen). Während die Korrelation zwischen Bilderwortschatz und Wortschatz mit 0,50 immerhin zu den höchsten erhaltenen Werten zählt, liegt der Zusammenhang zwischen den Tests Figurengruppen und Wortgruppen lediglich bei 0,36. Besonders der zuletzt genannte Wert ist auffällig, weil diese Tests denselben Faktor (Reasoning) messen sollten. Beide Tests korrelieren mit 0,44 bzw. 0,47 höher mit dem Zahlentest.

Die Interkorrelationen der PMA 13 - 14 liegen zwischen 0,06 (Wortschatz und Wahrnehmungsschnelligkeit) und 0,47 (Buchstabenreihen und Wahrnehmungsschnelligkeit). Auch hier ergeben sich die höchsten Koeffizienten zwischen analogen Tests zu den beiden Untersuchungsterminen 1964 und 1968. So korrelieren Wahrnehmungsschnelligkeit 1964 mit Wahrnehmungsschnelligkeit 1968 zu 0,63, Wortschatz 1964 mit Wortschatz 1968 zu 0,56 und Wortgruppen 1964 mit Wortgruppen 1968 zu 0,47.

ac) Real- und Oberschüler, N = 112

Unter der Diagonalen der Tabelle 14 sind die Interkorrelationen der Testrohwerte von 15 kognitiven Variablen für die Gruppe der Real- und Oberschüler eingetragen.

Die Hauptergebnisse:

Im Gegensatz zu den HS korrelieren hier mit dem Schulfreitest von KERN die PMA-Untertests Raumerfassung (1964) 0,30, Wahrnehmungsschnelligkeit (1964) 0,37, Figurenvergleich (1968) 0,32 und Wahrnehmungsschnelligkeit (1968) 0,35 doch beträchtlich höher als die meisten übrigen PMA-Untertests. Es werden mit diesen PMA-Untertests offenbar Fähigkeiten angesprochen, die beim Erlernen des Lesens und Schreibens von Bedeutung sind. Mit dem MÜNCHNER-Schulfreitest korrelieren außer dem Bilderwortschatz, den beiden Zahlentests und dem Wortschatz 1968 alle übrigen Untertests zwischen 0,30-0,42. Der MÜNCHNER-Schulfreitest scheint bei diesen Schülern eher Fähigkeiten der allgemeinen Intelligenz herzufordern.

Die Interkorrelationen der PMA 9 - 10 liegen zwischen 0,03-0,62 und die der PMA 13 - 14 zwischen -0,02 bis 0,47. Erwartungsgemäß erzielen jeweils die Tests, die gleiche Fähigkeit messen sollen, die höchsten Werte.

Ähnliche Relationen finden sich auch beim Vergleich der beiden Testbatterien: Wortschatz 1964 und Wortschatz 1968 korrelieren mit 0,70 resp. 49% gemeinsamer Varianz am höchsten. Nur unbedeutend niedriger ist mit 0,66 die Korrelation zwischen Bilderwortschatz 1964 und Wortschatz 1968. Die Koeffizienten der vier

Tests, die 1964 bzw. 1968 den Faktor "Reasoning" messen sollten, korrelieren zwischen 0,43-0,53. Raumerfassung (1964) und Figurenvergleich (1968), die den Faktor "Space" messen sollen, korrelieren lediglich mit 0,39 resp. 15% gemeinsamer Varianz. Dagegen korrelieren die beiden Tests Wahrnehmungsschnelligkeit mit 0,62 resp. 38% gemeinsamer Varianz relativ hoch.

Vergleicht man die Korrelationskoeffizienten der Gesamtstichprobe mit denen der HS und/oder denen der RS+OS so ist nicht zu übersehen, daß der größte Teil der Koeffizienten der beiden Teilstichproben erheblich niedriger ausgefallen ist. Dies wird in der Regel erwartet und traditionell mit der Streuungsreduktion der Variablen in den Teilstichproben erklärt. Wir halten eine derartige Erklärung für fragwürdig. Sie wäre nur dann zu erwarten, wenn die fraglichen Merkmale selbst nicht über das Ganze - für die Teilstichproben relevante - Kontinuum streuten, also z.B. aus der Teilstichprobe Gymnasiasten nur die Intelligentesten 10% der Schüler in der Stichprobe wären. Dies wiederum wäre eine Verletzung der Forderung nach Stichprobenrepräsentativität.

Um dies zu erläutern, seien zunächst die Modellannahmen dargestellt:

Gesetzt den Fall, man habe Tests, die hinsichtlich z.B. der Itemselektion den theoretischen Anforderungen für eine bestimmte Altersklasse genügten. Wenn nun diese Tests z.B. für 14jährige Schülerinnen und Schüler der Bundesrepublik geeicht werden sollen, wird dazu eine repräsentative Stichprobe der Population benötigt. Wenn ferner zu den relevanten Stichprobenmerkmalen auch die besuchte Schulart, z.B. SS, HS, RS, und OS, gehört, so verlangen die Modellannahmen folgendes:

- (1) Die Meßwerte verteilen sich in der Eichstichprobe um einen Mittelwert mit bestimmter Streuung normal;
- (2) Das gleiche gilt auch für Teilstichproben, die bestimmte Merkmale aufweisen, z.B. Besuch einer bestimmten Schulart. Auch für derartige Teilstichproben wird die Normalverteilung der Meßwerte postuliert, wobei sich der jeweilige Mittelwert im allgemeinen nach rechts oder links vom Gesamtstichprobenmittelwert verschiebt und die Streuung gegenüber derjenigen der Gesamtstichprobe geringer ist.

Das Postulat unter (2) impliziert jedoch nicht eine im Vergleich zur Gesamtstichprobe reduzierten Korrelation der (zwei) Merkmale in den Teilstichproben.

Vielmehr bedarf es für das z.B. um 50% reduzierte Produkt der beiden Streuungen in den Teilstichproben auch nur einer um - im Beispiel - 50% geringeren Kovarianz, um Korrelationskoeffizienten von derselben Höhe wie in der Gesamtstichprobe zu erhalten. Tatsächlich sind aber - wie auch bei uns - die Korrelationen in den Teilstichproben meistens geringer als in der Gesamtstichprobe. Wenn angenommen werden darf, daß weder die Tests in ihren Konstruktionsmerkmalen noch die Durchführungsbedingungen jeweils spezifische Einflüsse auf das Zustandekommen der Lösungen bei den Probanden in den verschiedenen Teilstichproben haben, bleiben u.E. nur zwei Erklärungsmöglichkeiten:

- Die Teilstichproben sind derart selegiert, daß die Kovarianz der zu vergleichenden Merkmale überproportional reduziert ist, im Verhältnis zum Produkt der Streuungen in den Teilstichproben der korrelierten Variablen.
- Die zu vergleichenden Variablen messen in den Teilstichproben (zumindest teilweise) etwas qualitativ anderes als die Korrelationen in der Gesamtstichprobe nahegelegt haben mögen.

In Wirklichkeit dürften aber alle genannten Einflußgrößen (Tests, situative Bedingungen, Stichprobe(n)) unsystematisch zusammenwirken.

b) Interkorrelationen der USW und GSW der Standardwerte der Tests

Um jedoch einen Überblick über die Verteilungskennwerte der Leistungsvariablen zu erhalten, haben wir zunächst die Mittelwerte und Streuungen von zehn (kumulativen) Werten für fünf Substichproben in Tabelle 15 zusammengestellt. Wie zu erwarten, steigen die Mittelwerte in allen Fällen linear an, von den SS über HS, RS zu den OS. Anders dagegen bei den Streuungen; mit Ausnahme des KERN haben die SS immer geringste Streuung. Weiterhin fällt auf, daß die HS nur im "Faktor" Lesefrei 9-10 die größte Streuung aufweisen, während dies bei der Gruppe der RS+OS in sieben Fällen gegeben ist. Dies deutet darauf hin, daß die Gruppe der weiterführenden Schüler nicht besonders homogen ist.

ba) Gesamtstichprobe, N = 224

Tabelle 16 enthält unter der Diagonalen die Interkorrelationen der Standardwerte für die Gesamtwerte der beiden Schulreifetests und der PMA 9-10 bzw. PMA 13-14, sowie der Tests, die die Faktoren "Verbal", "Reasoning" und "Lesefrei" zu den verschiedenen Untersuchungszeitpunkten erfassen sollen.

Die Hauptergebnisse:

Der Zusammenhang zwischen "KERN" und den GSW 9-10 bzw. GSW 13-14 ist mit 0,36 bzw. 0,33, resp. ca. 12% gemeinsamer Varianz nur noch gering. Anders da-

Tabelle 15 : Arithmetische Mittelwerte (M) und Streuungen (s) der Standardwerte für SS, HS, RS, OS und RS+OS

Normmittelwert = 50

Test		SS N=12	HS N=100	RS N=50	OS N=62	RS+OS N=112
Ges.-KERN	M	37,75	47,86	51,90	53,19	52,62
	s	9,12	9,06	7,98	7,52	7,72
Ges.-MÜNCHNER	M	36,67	46,83	51,76	56,29	54,27
	s	6,92	8,39	8,98	9,04	9,25
Verbal 9-10	M	34,00	45,17	52,48	58,76	55,96
	s	6,30	7,44	7,01	7,71	8,01
Reasoning 9-10	M	33,33	47,11	52,06	56,45	54,49
	s	7,27	7,92	8,50	8,72	8,86
Lesefrei 9-10	M	31,42	46,36	52,16	57,77	55,27
	s	5,70	8,08	7,31	7,24	7,76
GSW 9-10	M	31,17	45,73	52,28	58,74	55,86
	s	4,60	7,61	6,59	7,23	7,64
Verbal 13-14	M	31,75	44,78	53,68	58,90	56,57
	s	3,77	6,66	6,50	7,29	7,39
Reasoning 13-14	M	33,33	45,23	52,54	58,63	55,91
	s	5,87	7,46	6,85	7,57	7,84
Lesefrei(3)13-14	M	31,50	48,20	51,96	54,97	53,63
	s	5,00	7,86	9,40	9,12	9,32
GSW 13-14	M	30,33	45,70	53,02	58,27	55,93
	s	4,05	6,89	7,21	7,49	7,79

Tabelle 16 : Interkorrelationen (SPEARMAN) von Standardwerten nach McCALL, über Altershalbjahre, der beiden Schulreifetests 1960 (Tests 1+2), der USW und GSW der PMA 9-10 (Tests 3-6) und der SW, USW und GSW der PMA 13-14 (Tests 7-10)

Über der Diagonalen: Real- und Oberschüler, N = 112
Unter der Diagonalen: Gesamtstichprobe, N = 224

	1	2	3	4	5	6	7	8	9	10
1	—	.49	.12	.19	.29	.28	.08	.21	.32	.32
2	.61	—	.31	.43	.53	.55	.14	.43	.41	.47
3	.27	.47	—	.42	.59	.75	.76	.52	.24	.54
4	.27	.50	.55	—	.55	.72	.40	.63	.33	.56
5	.36	.58	.73	.74	—	.93	.47	.53	.57	.68
6	.36	.60	.81	.82	.97	—	.61	.65	.50	.73
7	.26	.45	.84	.60	.67	.76	—	.49	.09	.48
8	.28	.51	.66	.71	.73	.79	.72	—	.38	.76
9	.30	.43	.42	.51	.68	.64	.38	.56	—	.83
10	.33	.53	.69	.69	.82	.85	.73	.89	.84	—

1 = Gesamtwert-KERN

2 = Gesamtwert-MÜNCHNER

3 = USW-Verbal 9-10

4 = USW-Reasoning 9-10

5 = USW-Lesefrei 9-10

6 = GSW-PMA 9-10

7 = SW-Verbal 13-14

8 = USW-Reasoning 13-14

9 = USW-Lesefrei(3) 13-14

10 = GSW-PMA 13-14

Signifikanzgrenzen der Korrelationskoeffizienten:

RS+OS: $r > .20$ $p < 5\%$; $r > .26$ $p < 1\%$

Gesamt-N: $r > .14$ $p < 5\%$; $r > .18$ $p < 1\%$

gegen ist dies zwischen "MÜNCHNER" und den GSW 9-10 bzw. GSW 13-14 mit 0,60 und 0,53, resp. 36% bzw. 28% gemeinsamer Varianz. Diese Koeffizienten erscheinen relativ stabil über den Zeitraum von 4 bzw. 8 Jahren und kommen in ihrer Höhe als Maße für Schulreife und Intelligenz schon in die Nähe von Koeffizienten, die z.B. McCALL (nach JENSEN 1980, S. 279) für die zeitliche Stabilität von Intelligenzgrößen dieser Altersstufen gefunden hat. Die Korrelationen der "Faktoren"-Tests mit dem zugehörigen Gesamtstandardwert liegen für beide Testbatterien z.B. höher als die analogen Werte in der Handanweisung des Prüfsystems für Schul- und Bildungsberatung (PSB, HORN 1969), das ebenfalls nach der Faktorentheorie THURSTONE's entwickelt wurde. Die (zeitliche) Stabilität analoger Werte beider PMA-Tests kann als zufriedenstellend angesehen werden, wobei die Stabilität der Gesamtintelligenz mit 0,85 resp. 72% gemeinsamer Varianz und die Verbalisierungsfähigkeit mit 0,84 resp. 71% gemeinsamer Varianz offensichtlich gleichermaßen deutlich sind.

bb) Hauptschüler, $N = 100$; Sonderschüler, $N = 12$
Die Tabelle 18 enthält über der Diagonalen die Interkorrelationen der Standardwerte für die SS und unter der Diagonalen diejenigen der HS.

Bei den Schulreifetests fällt zunächst in der Gruppe der SS die negative Korrelation von $-0,39$ (wenn- gleich nicht signifikant) des "KERN" mit dem Reasoning-Faktor (1964) auf. Dies würde bedeuten, daß ein Schüler um so schlechter im Schulreife- test abschnitte je besser sein Ergebnis im schluß- folgernden Denken sei, und umgekehrt. Eine zutreffen- de Erklärung für diesen Wert erscheint schwierig, zumal 11 der 12 SS von den bei der Einschulung um ein Jahr vom Schulbesuch zurückgestellten Schülern gehören. Plausibel ist, daß der "KERN" bei über- alterten Schülern Schulreifemerkmale nicht mehr erfaßt, da die Schüler etwa die zu prüfende Gliede- rungsfähigkeit inzwischen beherrschen. Dieses Argument würde dadurch gestützt, daß die SS im "KERN" die größte Streuung ($s = 9,12$) aller schul- artspezifischen Gruppierungen haben. Außerdem streuen die SS-Ergebnisse im "KERN" weiter als in allen anderen Standardwerten (vgl. Tabelle 15).

Schon bei der Konstruktion bzw. Auswahl unserer Tests, aber auch bei den weiter beispielhaft angesprochenen Intelligenztests, die ja zu den besten gerechnet werden, scheint es u.E. nicht möglich gewesen zu sein, repräsentative Stichproben zu ziehen, die ja bereits für die Itemselektion notwendig sind. Mängel, die hier in Kauf genommen werden (müssen), sind im weiteren Verlauf der Testkonstruktion und Eichung kaum noch zu beheben.

Die bisherigen Ausführungen sollen mit einigen Beispielen belegt werden:

Die Korrelationskoeffizienten zwischen dem Gesamtwert-KERN einerseits und Raumerfassung (1964), Wahrnehmungsschnelligkeit (1964 und 1968) andererseits sind bei HS und RS+OS signifikant verschieden. Das gleiche gilt für den MÜNCHNER-Gesamtwert mit Figurengruppen bzw. Figurenvergleich. Und schließlich korreliert der Test Zahlen (1964) in der Teilstichprobe HS signifikant höher mit den Tests Figurengruppen und Wortgruppen als in der Teilstichprobe RS+OS.

Die Korrelation zwischen dem Test Zahlen (1964) und Figurengruppen ist bei den HS numerisch größer (nicht signifikant) als in der Gesamtstichprobe. In der Teilstichprobe RS+OS gilt Ähnliches für die Korrelationen zwischen Gesamtwert-KERN und Wahrnehmungsschnelligkeit (1964) sowie MÜNCHNER-Gesamtwert und Figurenvergleich

gegenüber der Gesamtstichprobe. In allen Fällen sind die Streuungsprodukte in den Teilstichproben geringer als in der Gesamtstichprobe.

Um die Bedeutung der nachfolgend dargestellten Korrelationskoeffizienten besser abschätzen zu können, seien für die Schüler der verschiedenen Schularten, bezogen auf die jeweilige Stichprobengröße, die Mindestgrößen der Korrelationskoeffizienten angegeben, die erreicht bzw. überschritten werden müssen, um auf dem 5%- oder 1%-Niveau signifikant zu sein. Tabelle 17 enthält die Mindestwerte.

Tabelle 17 : Mindestkorrelationskoeffizienten für die Signifikanzniveaus (p) 5% und 1%

Stichprobe		$p < 5\%$	$p < 1\%$
SS =	12	.58	.71
HS =	100	.20	.26
RS =	50	.28	.36
OS =	62	.25	.33
RS+OS =	112	.18	.23
Gesamt-N =	224	.13	.18

Tabelle 18 : Interkorrelationen (SPEARMAN) von Standardwerten nach McCALL, über Altershalbjahre, der beiden Schulreifetests 1960 (Tests 1+2), der USW und GSW der PMA 9-10 (Tests 3-6) und der SW, USW und GSW der PMA 13-14 (Tests 7-10)

Über der Diagonalen: Sonderschüler, N = 12

Unter der Diagonalen: Hauptschüler, N = 100

	1	2	3	4	5	6	7	8	9	10
1	—	.76	.02	-.39	-.07	.01	.39	.11	-.03	.14
2	.52	—	.35	-.17	.16	.31	.45	.17	.14	.39
3	-.06	.16	—	.34	.76	.81	.46	.39	.80	.88
4	-.01	.24	.29	—	.75	.67	.16	.51	.11	.28
5	.00	.28	.52	.70	—	.95	.31	.50	.57	.73
6	-.02	.31	.60	.80	.94	—	.33	.63	.63	.81
7	-.11	.22	.61	.40	.38	.48	—	.44	.42	.61
8	-.10	.15	.29	.49	.54	.59	.44	—	.22	.48
9	-.02	.18	.18	.38	.57	.51	.14	.46	—	.83
10	-.10	.23	.37	.53	.66	.68	.50	.84	.80	—

1 = Gesamtwert-KERN

2 = Gesamtwert-MÜNCHNER

3 = USW-Verbal 9 - 10

4 = USW-Reasoning 9 - 10

5 = USW-Lesefrei 9 - 10

6 = GSW-PMA 9 - 10

7 = SW-Verbal 13 - 14

8 = USW-Reasoning 13 - 14

9 = USW-Lesefrei (3) 13 - 14

10 = GSW-PMA 13 - 14

Signifikanzgrenzen der Korrelationskoeffizienten:

SS = 12: $r > .58$ $p < 5\%$; $r > .71$ $p < 1\%$

HS = 100: $r > .20$ $p < 5\%$; $r > .26$ $p < 1\%$

Ferner darf nicht übersehen werden, daß die Gruppe SS nur aus 12 Schülern besteht, so daß bereits geringfügige erwartungswidrige Resultate die negative Korrelation bewirkt haben können.

Vergleicht man analoge Korrelationskoeffizienten der HS und SS so fällt auf, daß die Koeffizienten bei den SS in vielen Fällen erheblich höhere Werte erreichen. Dies, obgleich die Streuungen der SS der zu vergleichenden Variablen oft erheblich unter denen der HS liegen. Dies gilt auch für viele analoge Koeffizienten der HS, RS, OS, RS+OS und auch der Gesamtstichprobe. Derartige Ergebnisse dürften gar nicht vorkommen, wenn die Streuungsreduktion allein für die häufig zu beobachtende Tatsache vermindelter Korrelationen bei Unterstichproben ursächlich wäre.

Was die Stabilitätswerte über 4 Jahre bei den SS und HS betrifft, so determinieren die "Faktoren" Verbal, Reasoning und Lesefrei in beiden Stichproben jeweils weniger als 50% der Varianz. Für die HS gilt dies auch für die Gesamtintelligenz.

bc) Real- und Oberschüler, N = 112

Die Tabelle 19 enthält über der Diagonalen die Interkorrelationen der Standardwerte für die RS und unter der Diagonalen diejenigen der OS.

Tabelle 19 : Interkorrelationen (SPEARMAN) von Standardwerten nach McCALL, über Altershalbjahre, der beiden Schulreifetests 1960 (Tests 1+2), der USW und GSW der PMA 9 - 10 (Tests 3-6) und der SW, USW und GSW der PMA 13-14 (Tests 7-10)

Über der Diagonalen: Realschüler, N = 50
Unter der Diagonalen: Oberschüler, N = 62

	1	2	3	4	5	6	7	8	9	10
1	—	.45	.14	.29	.30	.29	.06	.28	.29	.35
2	.50	—	.30	.44	.56	.55	-.02	.47	.34	.42
3	.07	.21	—	.37	.60	.70	.71	.40	.22	.45
4	.10	.39	.40	—	.59	.73	.34	.59	.24	.49
5	.31	.46	.49	.49	—	.95	.44	.60	.49	.68
6	.27	.48	.70	.70	.90	—	.53	.67	.42	.67
7	.07	.16	.76	.38	.40	.59	—	.42	.03	.40
8	.15	.34	.45	.63	.35	.55	.41	—	.31	.73
9	.34	.43	.17	.36	.58	.54	.04	.37	—	.82
10	.27	.44	.50	.60	.61	.73	.43	.76	.83	—

1 = Gesamtwert-KERN

2 = Gesamtwert-MÜNCHNER

3 = USW-Verbal 9 - 10

4 = USW-Reasoning 9 - 10

5 = USW-Lesefrei 9 - 10

6 = GSW-PMA 9 - 10

7 = SW-Verbal 13 - 14

8 = USW-Reasoning 13 - 14

9 = USW-Lesefrei (3) 13 - 14

10 = GSW-PMA 13 - 14

Signifikanzgrenzen der Korrelationskoeffizienten:

RS = 50: $r > .28$ $p < 5\%$; $r > .36$ $p < 1\%$

OS = 62: $r > .25$ $p < 5\%$; $r > .33$ $p < 1\%$

Die beiden Korrelationsmatrizen sind auffallend ähnlich hinsichtlich der Höhe analoger Werte in den beiden Substichproben z.B. die Schulreifetests mit den GSW's, die Interkorrelationen der PMA 9-10 bzw. PMA 13-14, aber auch unter dem Aspekt der Stabilität, d.h. analoger Werte beider PMA-Batterien.

Die Zusammenfassung der Real- und Oberschüler zur Gruppe Schüler weiterführender Schulen (RS+OS) bestätigt den Befund (vgl. Tabelle 16). Obgleich mit Ausnahme von zwei Fällen (RS: Ges.-KERN und Lesefrei(3) 13-14) alle Streuungen in der Gruppe RS+OS numerisch größer sind als in den Einzelgruppen RS und OS, sind doch eine beträchtliche Anzahl von Korrelationskoeffizienten in den beiden Stichproben mit reduzierten Streuungen numerisch (nicht signifikant) größer als bei RS+OS. Dieser Hinweis soll noch einmal unterstreichen, daß die Streuung nicht das alleinige Kriterium, ja nicht einmal das wichtigste für die Höhe der Korrelation zweier Variablen ist, sondern die Kovarianz der Variablen.

An diesem Beispiel können nochmals die beiden prinzipiell verschiedenen Vorgehensweisen demonstriert werden, wie Streuungs- und Kovarianzreduktion und der daraus resultierende mögliche Einfluß auf die Korrelationen zustande kommen können:

- (1) In unserem Beispiel kommen Streuungs- und Kovarianzreduktion dadurch zustande, daß die Ausgangsstichprobe nach einem demographischen Merkmal, hier: Schulart, in Untergruppen aufgestellt wird. Dies reduziert zwar im allgemeinen die Gesamt-Varianz in den Teilstichproben, muß aber die Kovarianz nicht überproportional reduzieren, so daß ein Einfluß auf die Höhe der Korrelationen nicht notwendigerweise gegeben ist.
- (2) Die zweite Möglichkeit, Streuung und/oder Kovarianz durch selektive Maßnahmen zu beeinflussen, kann dadurch geschehen, daß die zu korrelierenden Merkmale selbst als Selektionskriterium verwendet werden; wenn also z.B. aus unserer Stichprobe der HS nur diejenigen Probanden ausgelesen würden, deren Gesamtintelligenz (GSW) im PMA 9-10 und PMA 13-14 innerhalb genau definierter Grenzen liege.

KALVERAM (1969) hat dies für den Fall untersucht, daß eine Teilstichprobe (aus einer größeren) gebildet wird, mit Probanden,

"deren Testwerte die Bedingungen $x_1 + x_2 \sim 100$ erfüllen. Berechnet man nun in der kleineren Stichprobe ... den korrelativen Zusammenhang zwischen Test 1 und Test 2, so wird man einen Koeffizienten von nahezu -1 erhalten. In der (kleineren) Stichprobe ... nämlich ergänzen sich, wie (die) Bedingung ... besagt, die beiden Testwerte jeder Versuchsperson zu 100. Wenn bei einem Probanden das Ergebnis des einen Tests etwa nach oben abweicht, so muß es beim anderen im gleichen Maße - sozusagen 'kompensatorisch' - nach unten abweichen $x_1 \sim 100 - x_2$ Wählt man die kleinere Stichprobe ... aus der großen nach dem weniger strengen Kriterium $90 \leq x_1 + x_2 \leq 110$ aus, wird man natürlich keinen

Korrelationskoeffizienten von -1 erwarten; aber es dürfte einleuchten, daß auch hier zwischen den Testergebnissen eine negative kompensatorische Kovarianz auftreten muß (Hervorhebung vom Verf.), da die Variabilität der Summen-Noten eingeschränkt ist."

Hierbei bedeuten:

x_1 = Testwert im Test 1

x_2 = Testwert im Test 2

Daß bei Stichprobenselektionen, die die Bedingungen KALVERAM's erfüllen, immer negative Kovarianzen und damit auch negative Korrelationen auftreten, bedarf der Ergänzung. Wie unmittelbar einsichtig, kann eine selektive Stichprobe auch dadurch zustande kommen, daß die zuerst genannte Bedingung mit $x_1 = x_2$ und $x_1 + x_2 = 100$ erfüllt wird, d.h. wenn alle x_2 den gleichen Wert (hier: 50) haben. In diesem Fall sind aber Kovarianz und Korrelation gleich Null. Auch eine positive Kovarianz und damit eine positive Korrelation ist zu erzielen, wenn die genannte Bedingung $x_1 + x_2 \sim 100$ mit $x_1 = x_2$ paarweise erfüllt wird. Wir haben aus unserer Gruppe HS = 100 diejenigen Probanden herausgesucht, deren Ergebnis-Addition im Gesamtstandardwert der PMA 9-10 und PMA 13-14 möglichst nahe 100 ist. Dazu haben wir die Zusatzbedingungen formuliert, daß in der Summe der Werte 96 nicht unterschritten und 104 nicht überschritten werden sollte. Außerdem sollten alle Meßwerte die Forderung $47 \leq x \leq 53$ erfüllen. Dies ist u.E. eine stärkere Einschränkung der Varianzbreite der Meßwerte als die Bedingungen von KALVERAM sie fordern. Wir haben auf diese Weise 12 Probanden gefunden. Die Korrelation der beiden Meßwertreihen beträgt $r = +0,48$. Werden aus diesen 12 Probanden nur diejenigen ausgewählt, deren Meßwertsumme $99 \leq M \leq 101$ ist, so verbleiben 5 Probanden, deren Meßwerte zu $r = 0,93$ korrelieren. Die von KALVERAM postulierten Ergebnisse stellen sich notwendigerweise nur dann ein, wenn die Meßwertsummen aus Summanden resultieren, die eine möglichst große Differenz zueinander haben. Dieser Effekt tritt bei Meßwiederholungen um so deutlicher in Erscheinung, je stärker die sog. Regression zur Mitte ist.

Unsere Erörterungen und empirischen Befunde machen deutlich, daß die erwarteten Resultate sensu KALVERAM dann mit größter Wahrscheinlichkeit eintreten, wenn die weitere Auswahl der Probanden, die die gestellte Bedingung erfüllen, nur nach dem Zufallsprinzip erfolgt. Jede weitere "Manipulation" der Selektion kann - auch wenn dadurch die Varianz eingeschränkt wird - zu anderen Ergebnissen führen als von KALVERAM postuliert.

So haben wir unsere Gesamtstichprobe auf Grund des Gesamtstandardwertes in zwei Gruppen eingeteilt. Gruppe 1 GSW < 51, Gruppe 2 GSW > 50. Für diese Gruppen berechneten wir die Interkorrelationen der Standardwerte der PMA 9-10 und PMA 13-14. Die Koeffizienten sind - erwartungsgemäß - reduziert, in einigen Fällen auch negativ. Aber es gibt auch trotz der Einschränkung der Gesamtvariation bei analogen Tests beider PMA-Batterien - vor allem bei Tests mit verbaler Komponente - beachtliche positive Korrelationen in beiden Gruppen. So korrelieren z.B. die beiden Wortschatztests in Gruppe 1 zu 0,61 und in Gruppe 2 0,76. Die entsprechenden Werte für die beiden Wortgruppentests sind 0,46 bzw. 0,49.

Daraus folgern wir erneut die Unverzichtbarkeit möglichst präziser und vollständiger Diskription der Stichprobenselektion, da letztlich nur so nachvollzogen werden kann, wie oder warum Effekte zustande gekommen oder ausgeblieben sind. Dies sollte für Versuchspläne jeder Art gelten (vgl. auch LYKKEN 1968).

7.6

Validität

Die Validität liefert ein Maß für den Grad der Genauigkeit, mit dem ein Test das Merkmal mißt, das er zu messen vorgibt. Die von der AMERICAN PSYCHOLOGICAL ASSOCIATION (APA ⁴1974) vorgeschlagene Differenzierung unterscheidet drei Validitätsarten:

- criterion-related-validity, mit der Differenzierung nach predictive (= Vorhersagevalidität) und concurrent validity (= Übereinstimmungsvalidität)
- content validity (= Inhaltsvalidität)
- construct validity (= Konstruktvalidität).

Eine der Hauptabsichten der Längsschnittuntersuchung war es - wie bereits früher dargelegt wurde -

"spezifische Unterschiede zwischen Gruppen von erfolgreichen Kindern und von Schulversagern in den verschiedenen Persönlichkeitsbereichen und auf unterschiedlichen Leistungsgebieten ... zu analysieren und wenn möglich, Gruppen von Schulversagern zu unterscheiden" (KEMMLER 1967, S. 13).

Für den Fall, daß sich Schulversagen an bestimmte Merkmalskombinationen festmachen lassen sollte, bestand die Absicht, diese Kenntnisse zukünftig preventiv-prediktiv einzusetzen, mit dem Ziel, Schulversagen so weit wie möglich zu reduzieren.

Unter dem Aspekt der Chancengleichheit (vgl. AURIN 1968; JENCKS 1973; MÜLLER/MAYER 1976) sollte der Versuch unternommen werden, Hilfen für die Schullaufbahnberatung mit dem Ziel einer den individuellen Fähigkeiten und Interessen optimal entsprechenden allgemeinen Schulbildung bereitzustellen. Das heißt, die Untersuchung orientierte sich deutlich an den Problemen des Schulalltags. Diese Ausgangslage war folglich auch bestimmend für den Kontext der Validität: So ist anzumerken, daß etwa am Ende der Grundschulzeit (oder nachfolgender Orientierungsphase) für oder von jedem Schüler, der nicht eine Gesamtschule besuchen kann oder möchte die Entscheidung zu treffen ist, welcher Bildungsgang beschritten und welcher Abschluß angestrebt werden soll, unter der Maxime, einen möglichst hohen Schulabschluß bei Minimierung der Versagerquote zu erhalten.

Angewandt auf die Validität der zu erwartenden Befunde hieß dies folgendes:

Ausgehend davon, daß die kognitive Leistung(sfähigkeit), wie sie mit Hilfe von Intelligenztests gemessen wird, eine relevante Größe für Schulerfolg bzw. -versagen und Schulbildung sei, war es notwendig, Aussagen zur Validität der verwendeten Intelligenztests zu machen, zumal die verwendeten PMA-

Batterien (modifizierte) Übersetzungen darstellen, also keine auf deutsche Verhältnisse übertragbare Informationen vorlagen. KEMMLER (1967, 1976) hat dazu drei Wege beschritten:

- (1) In der Praxis des Schulalltags sind die Zensuren für den Bildungsgang der Schüler bedeutsam. Stellvertretend - weil besonders relevant - wurden daher die Noten in Deutsch und Mathematik und eine aus den beiden Noten resultierende Gesamtnote als Validitätskriterium zur Vorhersage des Schulerfolgs herangezogen. (Die Problematik der Zensuren wurde dabei nicht übersehen. Es sei hier nur auf INGENKAMP (1971) und SCHRÖTER (1977) verwiesen.)
- (2) Da es sich bei dieser Untersuchung um eine Längsschnittsstudie handelt, war der Aspekt der Übereinstimmungsvalidität der Tests bzw. Testwerte der verschiedenen Untersuchungszeitpunkte zu prüfen. Diesem Aspekt wurde zunächst dadurch Rechnung getragen, daß für die beiden Hauptuntersuchungen 1964 und 1968 jeweils analoge Testbatterien nach dem Modell von THURSTONE verwendet wurden. Die Übereinstimmungsvalidität ergibt sich somit aus den Interkorrelationen der analogen Tests beider Testbatterien. Außerdem wurde eine Kontrolluntersuchung mit N=116 HS, RS und OS durchgeführt. Von den Probanden wurden

die PMA 13-14 und das "Prüfsystem für Schul- und Bildungsberatung" (P-S-B von HORN 1969) bearbeitet. Das P-S-B wurde gewählt, weil es nach HORN auch gemäß dem Intelligenzmodell von THURSTONE entwickelt wurde und damals zu den besten deutschsprachigen Testverfahren gezählt wurde. Dies gilt offensichtlich noch immer. (vgl. HILTMANN 1977; WEISE 1975). Das P-S-B war also Kriterium für die PMA 13-14, d.h. die Interkorrelationen der PMA 13-13 mit dem P-S-B wurden als Maß für die Übereinstimmungsvalidität betrachtet.

- (3) Um das Konstrukt "Intelligenz" und dessen Teilkonstrukte, wie sie in den Faktoren THURSTONE's zum Ausdruck kommen, abschätzen zu können, wurden von KEMMLER auch Faktorenanalysen berechnet, jedoch nicht ohne Beachtung der vielfach diskutierten Problematik der Anwendung der Faktorenanalyse in diesem Kontext.

KEMMLER hat ihre Berechnungen zur Validität der Verfahren in ihren Büchern 1967, Seite 38f, und besonders ausführlich 1976, Seite 43-56, dargestellt. Eine zusammenfassende Wiedergabe der Ergebnisse müßte hier zu viele relevante Aspekte unberücksichtigt lassen.

Wir haben hier in Tabelle 20 lediglich einige Korrelationskoeffizienten der USW und GSW mit den gewichteten Gesamtnoten (aus Deutsch und Mathematik) für 1964 und 1968 von KEMMLER (1976, S. 43f) übernommen und um die Korrelationen mit dem Schulstatus 1968 erweitert. Beim Schulstatus handelt es sich um einen zweiziffrigen Index:

1. Ziffer von 1-4 für die vier Schularten,
2. Ziffer von 7-9 für die Klassenzugehörigkeit.

Lesebeispiele:

Den Schulstatus 17 (niedrigster Wert) erhielten Sonderschüler der Klasse 7, der niedrigsten Klasse zum Untersuchungszeitpunkt; den Schulstatus 49 erhielten Oberschüler der Klasse 9, der höchsten Klasse zum Untersuchungszeitpunkt 1968.

Etwas unerwartet ist für uns der relativ hohe Zusammenhang von Schulstatus und Testwerten. Er macht die große Bedeutung der kognitiven Variablen für den Schulerfolg deutlich. Ganz besonders zeigt sich dies bei den Faktoren "Verbal", "Reasoning" und den Gesamtstandardwerten, wobei "Reasoning" 1968 gegenüber 1964 noch deutlicher wird.

Für weitere Angaben müssen wir den Leser auf die angegebenen Quellen verweisen. Vergleicht man die Resultate mit dem in der Literatur vorfindbaren, so stimmen sie damit recht gut überein. Die Validi-

Tabelle 20 : Validitätskoeffizienten ^{+) der Testwerte mit den Kriterien Gesamtnote 64, Gesamtnote 68 und Schulstatus 68 über die Stichprobe der 1968 noch erreichten Schüler}

		Gewichtete Noten		
Tests		Gesamtnote 64	Gesamtnote 68	Schulstatus 68
Gesamtwert-KERN		.30	.31	.31
Gesamtwert-MÜNCHNER		.46	.41	.50
Verbal	9-10	.69	.58	.71
Reasoning	9-10	.62	.53	.57
Lesefrei	9-10	.71	.63	.66
GSW	9-10	.76	.67	.73
Verbal	13-14	.74	.63	.73
Reasoning	13-14	.74	.66	.70
Lesefrei(3)	13-14	.42	.56	.48
GSW	13-14	.69	.71	.71

Alle Koeffizienten sind auf dem 0,1%-Niveau signifikant.

^{+) Die Vorzeichen der Koeffizienten wurden so geändert, daß nicht der rein statistische, sondern der psychologisch relevante Zusammenhang deutlich wird.}

tätsergebnisse der Kontrolluntersuchung werden wir im Gesamtzusammenhang dieser Untersuchung, Seite 289ff, darstellen.

Wenn man die Konstruktion und den Einsatz psychologischer Tests als eine Technik auffaßt, so gilt für diese wie für jede andere Technik - soweit wir sehen -, daß es weitgehend vom Verwender abhängt, ob der Einsatz zum Vor- oder Nachteil der Betroffenen erfolgt (vgl. ANASTASI 1966, S. 15; EBEL 1963). Daraus folgern wir zweierlei:

- a) Die Testkonstrukteure müßten unmißverständlicher als derzeit erkennbar den Verwendungsbereich ihrer Instrumente angeben. Dadurch würde eine (fach-)öffentliche Kontrolle möglich und eine beliebige Verwendung zumindest potentiell eingeschränkt.
- b) Die Diskussionen über das Für und Wider der Verwendung psychologischer Tests sollten sich immer auf konkrete Sachverhalte beziehen, und sie müßten immer auch die Frage der Alternativen einschließen.

Globale Kritiken, zumal wenn sie zu dem Schluß kommen, daß der Einsatz von Tests eigentlich unverantwortlich sei, führen dann nicht recht weiter,

wenn keine Alternativen sichtbar werden. Ein Beispiel soll das verdeutlichen: REXILIUS (1976, S. 212f) kommt in seiner harschen Kritik der Intelligenz(-messung) zu folgendem Schluß:

"Alle Überlegungen und Fakten zusammengefaßt, ist das Ergebnis niederschmetternd - vor allem für den Getesteten, der beurteilt, über den entschieden werden soll. ... In Wirklichkeit schreiben sie (die Institutionen des bürgerlichen Staates) ihn, mehr oder weniger mobil fest und verstecken dieses beabsichtigte Resultat noch hinter der Illusion, Tests mit allen ihren sie zur Sinnlosigkeit verurteilenden Fragwürdigkeiten wären das richtige Mittel, dem Glück beizukommen."

Ein Ausweg wird darin gesehen,

"daß Informationen über Alter, Beruf und Erziehung, über die berufliche Vergangenheit bzw. allgemein sozial-statistische Daten und Informationen ... weitaus bessere, sicherere Vorhersagen gestatteten und zu brauchbareren Urteilen führten als Tests. ..."

Für Psychologen sollte es allerdings selbstverständlich sein, die Biographie der Probanden oder Klienten bei Entscheidungen immer miteinzubeziehen, insofern zeigt REXILIUS hiermit auch keinen Ausweg aus den objektiv bestehenden Schwierigkeiten der psychologischen Praxis. Es verwundert dann auch nicht, wenn REXILIUS seine Kritik mit der Einsicht abschließt:

"Tests können auch positive Funktion haben, sinnvoll und vorteilhaft eingesetzt werden, wenn mit ihrer Hilfe 'Unterprivilegierten' Getesteten Chancen, Möglichkeiten eröffnet werden können, die ihnen sonst mit Sicherheit vorenthalten blieben."

Das heißt aber nichts anderes, als daß Tests sehr wohl zu einer sichereren und kontrollierbaren, objektiven Entscheidungsfindung beitragen können, wenngleich auf dem Hintergrund eines bestimmten Klassenstandpunktes. Ob dies jedoch das einzige tragbare Fundament ist, kann an dieser Stelle offen bleiben. Unseres Erachtens müßten die von REXILIUS explizit und/oder implizit geäußerten Befürchtungen zu minimieren sein, wenn Testanwendungen etwa unter Berücksichtigung der sozialen Konsequenzen, wie EBEL (1963) sie formuliert hat, und bei Beachtung der "Ethical Standards of Psychologist" der APA (1963) erfolgten.

Bei den bisherigen Darstellungen und Analysen ließ es sich aus Gründen der Vollständigkeit der eigenen Arbeit nicht vermeiden, Ergebnisse KEMMLER's zu wiederholen bzw. nochmals mit aufzunehmen. Dies gilt sowohl für den Untersuchungsplan und die einzelnen Datenerhebungen als auch für Aspekte der Stichprobendarstellung und zu den instrumentellen Eigenschaften der psychologischen Verfahren. Wir haben uns jedoch bemüht, unsere Darstellungen und Analysen zur Stichprobe und den Testkriterien so vorzunehmen, daß die uns besonders interessierenden Aspekte deutlich wurden; so bei der Stichprobenbeschreibung durch die multidimensionale Analyse der Probandenmerkmale und bei Testkriterien durch den Vergleich verschiedener Tests sowie durch Gegenüberstellung von Korrelationskoeffizienten verschiedener Stichprobenuntergruppen.

Bei den noch zu bearbeitenden Fragen ist eine Überschneidung mit Befunden KEMMLER's nicht gegeben. Diese Fragen zur Längsschnittstudie und die Kontrolluntersuchung waren bisher unbearbeitet.

Der besseren Übersicht wegen wollen wir die im folgenden zu bearbeitenden Fragestellungen hier wiederholen:

- Im nächsten Kapitel werden wir uns mit dem

Phänomen der "Regression zur Mitte" bei Meßwiederholungen beschäftigen. Wir haben uns die Frage gestellt, ob Zunahme oder Abnahme von Testleistungen bei Meßwiederholungen - wie in der Fachliteratur behauptet - stets Methodenartefakte sind, oder ob es sich dabei doch um echte, psychologisch relevante Veränderungen handle.

- Daran anschließend werden wir der Frage nachgehen, ob mit Hilfe der multiplen Diskriminanzanalyse für Variablen des kognitiven Bereichs die individuelle Schullaufbahnberatung verbessert werden könne.
- Im letzten Kapitel des ersten Teils dieser Arbeit werden wir einen Aspekt der Anlage-Umwelt-Kontroverse zur Intelligenzentwicklung aufgreifen, Die umfangreichen Daten über die Schüler der Längsschnittstudie legen die Bildung von Pseudozwillingspaaren nahe. Wir wollen herausfinden, wie hoch der Anteil der Umweltvarianz, gemessen durch die Korrelationen der Gesamtstandardwerte, dieser Paare sei. Zudem können wir die Varianzschätzung zu zwei verschiedenen Meßzeitpunkten berechnen, ein Aspekt, der in dieser Kontroverse unseres Wissens deswegen unberücksichtigt bleiben mußte, weil die dazu notwendigen Daten nicht zur Verfügung stehen. Ergänzend dazu werden wir

für einige Zwillingsgruppen aus der Fachliteratur homogenere Untergruppen bilden und dafür neue Korrelationskoeffizienten berechnen. Die Homogenisierung soll uns Aufschluß über das Zustandekommen der Korrelationskoeffizienten liefern. Von unseren Ergebnissen erwarten wir Anregungen für die Diskussion des herkömmlichen Modells der Varianzschätzung in der Anlage-Umwelt-Kontroverse.

- Den zweiten Teil unserer Arbeit bildet eine Kontrolluntersuchung. Hauptziel ist es, die Übereinstimmungsvalidität der PMA 13-14 der Längsschnittstudie zu ermitteln, d.h. diese Untersuchung wurde durchgeführt, um die PMA 13-14 mit geeichten Tests, dem "Prüfsystem für Schul- und Bildungsberatung" (P-S-B) und dem "Frankfurter Analogie-Test" (FAT 7-8), vergleichen zu können. Dabei werden wir die Gelegenheit nutzen, unsere Befunde aus der Längsschnittstudie hinsichtlich der Veränderungsmessung kognitiver Strukturen sowie der Verwendung der multiplen Diskriminanzanalyse als Hilfsmittel bei der Schullaufbahnberatung einer ersten Überprüfung zu unterziehen.

8.0

VERÄNDERUNGSMESSUNG KOGNITIVER MERKMALE
AN KORRELIERTEN STICHPROBEN

Vorbemerkungen:

Bei Veränderungsmessungen resp. Meßwiederholungen stößt man unvermeidlich auf das verteilungs-statistische Phänomen der Regression zur Mitte (vgl. TYLER ³1965, S. 376; KLAUER 1973; STAKE 1975, S. 289; WILLERMAN 1979, S. 35f). Mindestens zwei Aspekte sind dabei zu unterscheiden:

- a) Eltern-Kind-Relationen (Generationsaspekt),
- b) intra-individuelle Merkmalsveränderung bei wiederholter Messung.

Unter dem Begriff der "Konstanz des Intelligenzquotienten" haben sich bereits STERN (1925) und WEINERT (1925) mit dieser Frage beschäftigt!

In jüngster Zeit haben sich eine schier unübersehbar große Zahl von Publikationen mit dieser Problematik auseinandergesetzt (RENN 1974; MANDL 1975, dort werden auch die bekanntesten älteren Untersuchungen zitiert; GROMBACH 1975; HUSEN 1976; OVERTON 1978; HELLER/ROSEMAN/STEFFENS 1978; KÜHN/HESSE 1979; ROGAN/KESSELMAN/MENDOZA 1979; BUSE 1980; LABOUVIE 1980; RUDINDER 1981).

Der Generationsaspekt meint das Phänomen, daß sehr große Eltern im allgemeinen Kinder haben, deren Körpergröße kleiner ist als die eigene und zur Populationsmitte tendiert. Umgekehrt haben sehr kleine Eltern im allgemeinen größere Kinder, ebenfalls mit der Tendenz zur Populationsmitte.

Bei wiederholter Messung, sagen wir des IQ, erzielen Probanden, die bei der ersten Messung einen sehr niedrigen IQ erzielten, bei der Folgemessung einen höheren, mit der Tendenz zum wahren Wert. Analog dazu erzielen Probanden, die bei der ersten Messung einen sehr hohen IQ erzielten, in der zweiten Messung einen niedrigeren.

Dieser rein statistisch zu erwartende Sachverhalt und die daraus erwachsenden methodologischen Probleme haben zu Argumenten gegen derartige Designs geführt (SCHAIE 1979, S. 196ff), da auf diese Weise erzielte Effekte Artefakte darstellten und somit besonders, wenn sie aus pädagogischen oder entwicklungspsychologischen Untersuchungen resultierten, nicht interpretierbar seien.

Diese Schwierigkeit sieht auch TYLER (TYLER ³1965, S. 375f), wenn sie eine Untersuchung von SKEELS (1940) über den sehr positiven Einfluß (IQ-Zuwachs)

bei geistesschwachen (feeble-minded) Kleinkindern, der ihnen durch die Obhut von ebenfalls geistes- schwachen erwachsenen Heimbewohnern zuteil geworden war, beurteilt:

"Es erscheint jedoch unwahrscheinlich, daß man mit Bezug auf statistische Artefakte alle Steigerungen erklären kann, die Skeels berichtet" (TYLER 1965, S. 376.

Vor ähnlichen Problemen stehen auch viele Kontrahenten der Diskussion um die kompensatorische Erziehung in den USA (vgl. HELLMUTH 1970).

Worum geht es bei der Veränderungsmessung?

Nun, die Diskussion ist schon alt - wenngleich noch keine befriedigende Lösung in Sicht ist (CRONBACH/FURBY 1970, S. 68; THORNDIKE 1966, S. 121) - so daß wir es uns hier versagen können, allzu häufig Veröffentlichtes lediglich erneut zu rekapitulieren. Der Interessierte Leser findet Darstellungen in dem von HARRIS (1963) herausgegebenen Konferenzbericht von 1962, wo ausschließlich diese Thematik diskutiert wird. Deutschsprachige Publikationen finden sich von PETERMANN (1978) und HELMREICH (1977). Im IPN-Kiel hat im Oktober 1979 ein Seminar unter Leitung von ROST und SPADA mit dem Titel "Veränderungsmessung zur Diagnose und Prognose von Lerneffekten"

stattgefunden. Diese Symposien mit dem sehr eingegrenzten Gegenstand deuten auf den besonderen Stellenwert dieser Problematik hin.

Soviel sei aber doch gesagt: Bei Längsschnittuntersuchungen werden (an derselben Stichprobe) häufig gleiche Merkmale (z.B. Intelligenz, Lernfortschritt, etc.) wiederholt gemessen. Die Frage ist nun jeweils, worauf Veränderungen, sofern sie eingetreten sind, zurückzuführen seien. Handelt es sich um echte Effekte oder um Artefakte? Sind sie Folge des Treatments oder nur auf andere situative und/oder statistische Einflüsse zurückzuführen? Zu letzteren gehören:

- o Übungs- und Durchführungseffekte,
- o Stichprobenfehler,
 - Auswahl (Repräsentativität),
 - Volunteer-Subject-Bias
(= Verzerrung durch Freiwilligkeit)
 - Probanden-Retrieval
(= Wiederauffindung der Probanden)
 - Mortalität
(= Ausscheiden durch Tod)
- o Reliabilitätsprobleme
- o Regressionseffekte

HELMREICH (1977, S. 23) postulierte für Veränderungs-
messungen zwei notwendige Bedingungen:

"1. Es muß um eine Variable gehen, bei der die Mög-
lichkeit besteht, daß sich ihre Werte im Laufe der
Zeit quantitativ verändern.

2. Die Variable selbst soll dabei unverändert blei-
ben."

Die erste Bedingung mag hier als erfüllt angesehen
werden. Bei der zweiten liegen die Verhältnisse schon
anders: Woher wissen wir, daß die Intelligenz (unser
Gegenstand) sich zwischen den Messungen nicht quali-
tativ verändert hat? Wo wir doch nicht einmal wis-
sen, was Intelligenz ist. +)

GUILFORD (1972) hat erst kürzlich den Nachweis zu
führen versucht, daß sich die hier verwendete In-
telligenztheorie THURSTONE's nur teilweise als
Theorie einer Structure-of-Intellect Abilities eigne.

+) IM (1975) hat dazu einen Großteil der Literatur
in einer Arbeit mit dem Titel: "Die Entwicklung
eines europäischen Schlüsselwortes: Intelligenz
und seine Bedeutung in der Wissenschaftssprache"
aufgearbeitet. Danach sind mehr als 60 voneinan-
der abweichende Definitionen von Intelligenz be-
kannt. Um mit dem eigenen Standpunkt jedoch nicht
hinter dem Berg zu halten, wollen wir bekennen,
daß wir eine Definition von HOFSTÄTTER (1971, S.
196ff) - bei IM nicht enthalten - favorisieren:
Danach ist Intelligenz die Fähigkeit zur Auffindung
von Ordnung. Natürlich hat auch diese Definition
ihre Probleme. Sie hat aber u.E. unbestreitbar
einen Vorteil, den der Einfachheit. Sie scheint
uns auch in der Praxis des Alltags Personen nicht
zu diskriminieren. Vor allem aber faßt sie in
aller Kürze zusammen, worin denn in summa intelli-
gente Aktivitäten letztlich ihren Ausdruck finden.

Da jedoch auch GUILFORD's Strukturtheorie der Intelligenz keineswegs problemlos ist, läßt sich so auch unser Rekurs auf THURSTONE rechtfertigen.

Reliabilität, aber erst recht Validität sind bei allen Tests noch ein gutes Stück von ihrem Maximalwert entfernt. Akzeptieren wir aber das Postulat und wollen trotz der Unsicherheiten Veränderungsmessungen studieren, bleibt uns nur die Möglichkeit, die Bedingungen als hinreichend genau realisiert anzunehmen.

HELMREICH (1977, S. 64ff) vergleicht drei Modelle zur Feststellung individueller Differenzen

- (1) den Differenzwert D ,
- (2) den "wahren" Veränderungswert,
- (3) den Change-Quotienten als Ausweitung eines Maßes residualisierter Veränderung.

Der Autor stellt einen hohen korrelativen Zusammenhang im Ergebnis der drei Modelle fest und kommt zu dem Schluß:

"Wenn es nicht um das Veränderungsmaß eines Probanden selbst geht, und dieser z.B. mit der Veränderung eines anderen Probanden verglichen werden soll, sondern wenn weitere Auswertungen mit den Veränderungsdaten gemacht werden sollen, kann man wohl auf den am einfachsten zu errechnenden Differenzwert D zurückgreifen."

Aus diesem Grund und wegen der unmittelbaren Anschaulichkeit haben wir die Veränderungswerte nach dem Differenzwert-Modell D berechnet. Die Berechnung von D erfolgt nach der Formel

$$\underline{D = \text{Meßwert der 2. Messung} - \text{Meßwert der 1. Messung};}$$

Auf diese Weise gibt das Vorzeichen von D an, ob es sich um eine Zunahme oder Abnahme der Veränderung handelt.

8.1 Verteilungsstatistische Differenzen der Meß-
wiederholungen bei verschiedenen Tests=====

Hier soll zunächst untersucht werden, wie sich die Wiederholungsmeßwerte analoger Testleistungen bei Schülern verschiedener Schularten (SS, HS, RS, OS) auswirken, d.h. wie stark die Regression zur Mitte durch die Verteilungen anschaulich sichtbar wird.

Erwartungsmäß finden sich bei allen neun Differenz-Verteilungen in den Extrembereichen (größer $\pm$ 4 Punkte) jeweils nahezu übereinstimmende relative Häufigkeiten (jeweils letzte Spalte der Tabelle 21 bis 29). Diese Erwartung ist dadurch begründet, daß die Standardisierungen der Tests über die Untersuchungsstichprobe erfolgt sind. Einige Hinweise auf die relative Stabilität der verschiedenen Intelligenzfaktoren vermitteln jedoch die relativen Häufigkeiten im Mittelbereich der Differenz-Verteilungen (bis $\pm$ 4 Punkte). Danach erweist sich die Gesamtintelligenz (GSW) bei 64% (vgl. Tabelle 21) aller Schüler über die Altersspanne von 10 - 14 Jahren als sehr stabil. Mit 57% ist auch der Gesamtanteil der Probanden, deren "Verbale Intelligenz" in 4 Jahren nur um maximal 4 Punkte variiert, relativ hoch (vgl. Tabelle 22). Für die Untersuchung 1968 stand dabei nur ein Verbaltest zur Verfügung, dessen Reliabilität mit 0,76 jedoch nicht als besonders hoch eingestuft werden kann.

Es ist denkbar, daß zwei Tests (wie 1964) - mit einer höheren Gesamtreliabilität - den Wert von 57% überstiegen hätte.

Von den bei dieser Längsschnittuntersuchung getroffenen faktoriellen Differenzierungen ist die "Reasoning-Intelligenz" mit 42% Schüler im genannten Bereich die instabilste; oder anderes: Die Fähigkeit, Problem logisch zu lösen, Einsicht zu gewinnen, zu planen, etc. im Alter von 10 - 14 Jahren noch größerer Veränderung zugänglich (vgl. Tabelle 23).

Ebenso im Bereich "Nicht-verbaler Intelligenz"; hier findet sich mit 44% fast die gleiche Instabilität resp. Veränderbarkeit wie bei der Reasoning-Fähigkeit (vgl. Tabelle 24). Die Gesamtreliabilitäten der nicht-verbalen Testskores ($r_{tt10} = 0.92$; $r_{tt14} = 0.98$) sind dabei gleich hoch wie diejenigen der Gesamtstandardwerte ($r_{tt10} = 0.94$; $r_{tt14} = 0.95$).

Die Stabilität der Testleistungen sinkt erwartungsgemäß im Vergleich zwischen GSW 13-14 bzw. GSW 9-10 und den beiden Schulreifetests. Ein Vergleich des MÜNCHNERSchulreifetests mit den beiden GSW's liefert im mittleren Differenz-Bereich mit 39% bzw. 38% eine praktisch vollständige Übereinstimmung der Werte (vgl. Tabellen 25 und 26). Für den KERNSchulreifetest liegen die analogen Zahlen bei 31% bzw. 30% (vgl. Tabellen 27 und 28). Eine Erklärung

für diese numerisch gleichen Werte je Schulreife-
test mit den in vierjährigem Abstand ermittelten
GSW's kann aus den Daten nicht abgeleitet werden.

Waren die relativen Differenzen im mittleren Bereich
für die GSW's recht hoch, so finden sich im gleichen
Differenzbereich für den MÜNCHNER-/KERNschulreife-
test nur 40% der Stichprobe (vgl. Tabelle 29),
obgleich - wie bereits ausgeführt - beide Tests am
selben Tag durchgeführt wurden. Für diese geringe
Übereinstimmung der Ergebnisse sind verschiedene
Faktoren ursächlich; etwa: Übungeffekte, verschie-
dene Validität, etc. So beträgt die Interkorrelation
der beiden Tests über die Gesamtstichprobe ($N = 224$)
lediglich $r = 0,61$, entsprechend einer gemeinsamen
Varianz von nur 37%.

a) Veränderungen im Gesamtstandardwert: $GSW_{14} - GSW_{10}$

Die Differenzen zwischen den beiden Gesamtstandard-
werten $GSW_{13-14} - GSW_{9-10}$ wurden für alle 224
Schüler ermittelt, wobei die Vorzeichen die Rich-
tung der Veränderungen im zeitlichen Verlauf ange-
ben. Die Trennung der Schüler nach besuchter Schul-
art soll bildungsspezifische Veränderungen aufdecken.
Tabelle 21 zeigt nun, wie die Differenzen in den
einzelnen Gruppen kumulieren.

Tabelle 21 : Häufigkeiten der Standardwertdifferenzen
GSW₁₄ - GSW₁₀ nach Schulart getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+14			2						2	
+11 bis +13			2				1		3	
+ 8 " +10			2	18	4	28	2	18	8	19
+ 5 " + 7			12		10		8		30	
+ 2 " + 4	3		15		13		14		45	
+ 1 " - 1	3	100	30	64	6	58	12	63	51	64
- 2 " - 4	6		19		10		13		48	
- 5 " - 7			8		3		4		15	
- 8 " -10			7		3		4		14	
-11 " -13			3	18	0	14	3	19	6	17
-14 " -15					1		1		2	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Die Differenzen der SS bewegen sich lediglich im Bereich von +4 bis -4 Punkten mit einer relativen Tendenz zum Negativen.
- (2) Die größten absoluten positiven Veränderungen finden sich bei HS (14 Punkte), die größten absoluten negativen bei RS und OS (bis -15 Punkte).
- (3) Die Gruppe der RS weist die relativ größte Anzahl von Schülern auf, die sich um 5 und mehr Punkte verbessert haben (N=14; 28% von N=50), während die Zahl bei den HS und OS nur jeweils 18% beträgt.
- (4) Eine statistische Verteilungsprüfung (9 Feldertafel) über die HS, RS und OS in den 3 in der Tabelle angegebenen Differenzklassen ist nicht signifikant.

b) Veränderungen in den Verbalwerten:

USW-Verbal₁₄ - USW-Verbal₁₀

Hier und für die folgenden Darstellungen werden die Differenzen analog zu den GSW-Differenzen gebildet und ebenso statistische Verteilungsprüfungen vorgenommen. Tabelle 22 zeigt die Verteilung der Standardwertdifferenzen in den Verbaltests.

Tabelle 22 : Häufigkeiten der Standardwertdifferenzen
USW-Verbal₁₄ - USW-Verbal₁₀ nach Schulart getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+14 bis +16			1		1				2	
+11 " +13			2		1				3	
+8 " +10		17	5	21	2	30	6	21	13	23
+5 " +7	2		13		11		7		33	
+2 " +4	2		14		9		12		37	
+1 " -1	0	33	19	59	14	56	17	63	50	57
-2 " -4	2		26		5		10		43	
-5 " -7	5		11		4		7		27	
-8 " -10	0		6		2		2		10	
-11 " -13	1	50	2	20	1	14	0	16	4	19
-14 " -16			0		0		1		1	
-17 " -19			1		0		0		1	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Die Streubreite der Veränderungen ist unter den SS bei den Verbaltests größer als zwischen den GSW's. Allerdings bleibt auch hier die sehr starke negative Tendenz erhalten.
- (2) Die größten absoluten positiven Veränderungen (bis zu +16 Punkten) finden sich bei den HS und RS, die absolut größten negativen bei den HS (bis zu -19 Punkten).

- (3) Wie schon bei den GSW ist auch hier unter den RS mit 30% die relativ größte Anzahl von Schülern mit positiven Differenzen im Verbalbereich zu finden. Die relativ geringsten Quoten negativer Veränderungen zeigen die RS und OS. Anders formuliert: Die SS und HS zeigen die relativ größten negativen Veränderungen.
- (4) Die Verteilungsprüfung ist nicht signifikant von Null verschieden.

c) Veränderungen in den Reasoningwerten:

USW-Reasoning₁₄ - USW-Reasoning₁₀

Tabelle 23 zeigt die Verteilungen der von den Schülern in den Reasoningtests erzielten Differenzen.

Tabelle 23 : Häufigkeiten der Standardwertdifferenzen USW-Reasoning₁₄ - USW-Reasoning₁₀ nach Schulart getrennt

Differenz		SS		HS		RS		OS		Ges.	
		f	%	f	%	f	%	f	%	f	%
+17 bis	+19			1		0		0		1	
+14 "	+16			2		2		3		7	
+11 "	+13		33	2	23	1	32	4	37	7	29
+ 8 "	+10	2		8		6		6		22	
+ 5 "	+ 7	2		10		7		10		29	
+ 2 "	+ 4	0		11		6		11		28	
+ 1 "	- 1	3	42	12	38	9	44	10	48	34	42
- 2 "	- 4	2		15		7		9		33	
- 5 "	- 7	2		15		4		3		24	
- 8 "	-10	0		11		4		4		19	
-11 "	-13	1	25	8	39	2	24	1	15	12	28
-14 "	-16			4		1		0		5	
-17 "	-19			0		1		0		1	
-20 "	-22			0		0		1		1	
-23 "				1		0		0		1	
		12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Unter allen Testdifferenzen erzielten die SS hier ihre absolut größten positiven Veränderungen (bis zu +10 Punkte. Gleichzeitig gibt es aber auch einen SS, dessen negative Veränderung mehr als eine Standardeinheit (-13 Punkte) beträgt.
- (2) Die absolut größte positive (bis +19 Punkte) Veränderung wie auch die absolut größten negativen (bis -23 Punkte) finden sich bei den HS. Relativ betrachtet, ist zwischen den Werten der HS, RS und OS ein starkes Mißverhältnis - besonders bezogen auf HS und OS - festzustellen; während bei den OS 37% positive Veränderungen von mehr als 4 Standardwertpunkten aufweisen, liegt diese Zahl bei den HS bei nur 23%. Entsprechend sinken bei 39% der HS die Testwerte um mehr als -4 Punkte, während von den OS in diesem Bereich nur 15% liegen.
- (3) Die Gleichverteilungsprüfung der Testwertfrequenzen der HS, RS und OS (9 Felder) ist in der unter (2) beschriebenen Weise auf dem 5%-Niveau signifikant. Dabei haben auch die RS mit 32% noch einen erwartungswidrig positiven "Überhang".

d) Veränderungen in den nicht-verbalen Werten:
USW-Lesefrei₁₄ - USW-Lesefrei₁₀

Tabelle 24 zeigt die Verteilungen der von den Schülern in den nicht-verbalen Tests erreichten Differenzen, die sich im zeitlichen Abstand von 4 Jahren ergeben haben.

Tabelle 24: Häufigkeiten der Standardwertdifferenzen USW-Lesefrei₁₄ - USW-Lesefrei₁₀ nach Schulart getrennt

Differenz		SS		HS		RS		OS		Ges.	
		f	%	f	%	f	%	f	%	f	%
+17	bis +19			2		1		0		3	
+14	" +16			3		2		1		6	
+11	" +13		33	6	36	1	30	0	13	7	28
+8	" +10			9		4		3		16	
+5	" 7	4		16		7		4		31	
+2	" +4	0		18		6		8		32	
+1	" -1	2	42	15	47	10	38	12	45	39	44
-2	" -4	3		14		3		8		28	
-5	" -7	3		7		4		6		20	
-8	" -10			5		6		8		19	
-11	" -13		25	3	17	3	32	7	42	13	28
-14	" -16			2		3		3		8	
-17	" -19							0		0	
-20	" -22							1		1	
-23								1		1	
		12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Die Streubreite beträgt für die SS maximal plus/minus 7 Punkte. Sie liegt damit nur geringfügig oberhalb derjenigen, die sie bei den GSW's erreichen.
- (2) Die absolut größten positiven Veränderungen finden sich bei den HS und RS mit bis zu +19 Punkten, die absolut größten negativen erzielten die OS mit -23 Punkten.
- (3) Der Befund unter (2) bleibt auch bei relativer Betrachtung der Veränderungen in den Gruppen der HS, RS und OS erhalten. Während von den HS 36% positive Veränderungen von mehr als +4 Punkten erzielten, sind es bei den OS nur 13%. Umgekehrt ver-

schlechtern sich bei den HS die Testleistungen bei nur 17% um mehr als -4 Punkte, dagegen sind es bei den OS 42%. Die Veränderungen bei den RS sind in den angegebenen Bereichen etwa ausgeglichen.

- (4) Die Gleichverteilungsprüfung ist in dem unter (3) angegebenen Sinne auf dem 1%-Niveau signifikant.

e) Veränderungen zwischen Intelligenzleistungen und
Schulreifetestergebnissen GSW₁₀ - MÜNCHNER-schul-
reifetestgesamtwert

Derartige Vergleiche sind nicht unproblematisch, denn es ist nicht die Funktion von Schulreifetests, die allgemein Intelligenz in der Weise zu erfassen, wie dies Intelligenztests zu tun beanspruchen. Andererseits sind die Korrelationen zwischen den Tests über die Zeiträume von 4 bzw. 8 Jahren nicht so gering, daß ein Vergleich völlig sinnlos erschiene. Immerhin müssen besonders hier bei der Interpretation die Korrelationen beachtet werden. Tabelle 25 zeigt die Verteilungen der Differenzen, die von den Probanden in den o.a. Tests erzielt haben.

Tabelle 25: Häufigkeiten der Standardwertdifferenzen GSW - MÜNCHNERschulreifetestgesamtwert nach Schul-¹⁰art getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+29							1		1	
+26 bis +28							0		0	
+23 " +25							1		1	
+20 " +22							0		0	
+17 " +19			5	28	0	30	1	32	6	29
+14 " +16		8	2		2		2		6	
+11 " +13			5		3		5		13	
+ 8 " +10			4		5		3		12	
+ 5 " + 7	1		12		5		7		25	
+ 2 " + 4	2		14		10		16		42	
+ 1 " - 1	2	42	6	34	3	38	7	44	18	38
- 2 " - 4	1		14		6		4		25	
- 5 " - 7	0		12		11		8		31	
- 8 " -10	2		11		1		4		18	
-11 " -13	1	50	10	38	4	32	3	24	18	33
-14 " -16	3		2		0		0		5	
-17 " -19			1						1	
-20 " -22			0						0	
-23 " -25			0						0	
-26 " -28			1						1	
-29			1						1	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) 50% der SS zeigen negative Veränderungen von mehr als -4 bis -16 Punkten, und nur ein SS erzielt eine positive Veränderung von +5 Punkten.
- (2) Der absolut höchste Wert (-29 Punkte) sowie die relativ größte Zahl (38%) negativer Veränderungen findet sich bei den HS.
- (3) Dagegen zeigen die OS sowohl absolut (bis +29 Punkte) als auch relativ (32%) die stärksten positiven Veränderungen von mehr als +4 Punkten auf.
- (4) Die Verteilungsprüfung ist jedoch nicht bedeutsam von Null verschieden (9 Felder).

f) Veränderungen zwischen Intelligenztestleistungen und
Schulreifetestergebnissen: GSW₁₄ - MÜNCHNERschulreife-
testgesamtwert

Tabelle 26 zeigt die Verteilung der Differenzen zwischen
dem Gesamtstandardwert (GSW 13-14) und dem Gesamtwert im
MÜNCHNER Schulreifetest.

Tabelle 26: Häufigkeiten der Standardwertdifferenzen GSW₁₄ -
MÜNCHNERschulreifetestgesamtwert nach Schul-
art getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+23 bis +25			1						1	
+20 " +22			0		2		1		3	
+17 " +19			2		0		4		6	
+14 " +16			4	28	2	38	3	37	9	31
+11 " +13			4		2		3		9	
+ 8 " +10			5		7		6		18	
+ 5 " + 7			12		6		6		24	
+ 2 " + 4	2		8		4		5		19	
+ 1 " - 1	1	42	15	36	5	38	12	44	33	39
- 2 " - 2	2		13		10		10		35	
- 5 " - 7	2		13		2		3		20	
- 8 " -10	1		7		6		3		17	
-11 " -13	2	58	8	36	3	24	4	19	17	30
-14 " -16	2		3		0		2		7	
-17 " -19			1		1		0		2	
-20 " -22			2		0		0		2	
-23 " -33			2		0		0		2	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Für die SS ist die relative Anzahl negativer Veränderungen gegenüber den Ergebnissen unter e) auf 58% angestiegen.
- (2) In absoluten Werten zeigen die HS sowohl positiv (bis +25 Punkte) wie negativ (bis -33 Punkte) die größten Veränderungen.

- (3) 28% des HS, aber 37% der OS zeigen einen Anstieg von mehr als +4 Punkten. Umgekehrt fallen 36% der HS, aber nur 19% der OS um mehr als -4 Punkte ab.
- (4) Auch bei den RS verbessern sich im genannten Bereich 38%, während sich nur 24% verschlechtern.
- (5) Die Gleichverteilungsprüfung (9 Felder) weicht im Ergebnis deutlicher als unter e) von Null in der sich aus den Prozentangaben ergebenden Weise ab, ist aber nur auf dem 10%-Niveau signifikant.

g) Veränderungen zwischen Intelligenztestleistungen
und Schulreifetestergebnissen: $GSW_{10} - KERNschul-$
reifetestgesamtswert

Die Bildung von Differenzen zwischen GSW_{9-10} bzw. GSW_{13-14} und den Ergebnissen im KERNschulreifetest wird hier für die SS und HS problematisch, denn für diese beiden Gruppen ergeben sich keine signifikanten Zusammenhänge mehr zwischen den jeweils zu vergleichenden Tests. Ob und in welcher Weise die für diese Gruppen vorliegenden Verteilungen etwa zur Abschätzung von Regressionseffekten dennoch von Nutzen sein können, muß der Gesamtdiskussion der Ergebnisse vorbehalten bleiben. Tabelle 27 zeigt die Verteilungen der Differenzen zwischen dem Gesamtstandardwert GSW_{10} und dem Gesamtwert im KERNschulreifetest.

Tabelle 27: Häufigkeiten der Standardwertdifferenzen GSW₁₀ - KERNschulreifetestgesamtwert nach Schulart getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+32 bis +33			1		0		0		1	
+29 " +31			1		0		0		1	
+26 " +28			0		0		0		0	
+23 " +25			0		0		4		4	
+20 " +22		8	0	31	1	30	1	52	2	35
+17 " +19			0		0		4		4	
+14 " +16			7		1		4		12	
+11 " +13	1		7		5		2		15	
+ 8 " +10	0		6		4		6		16	
+ 5 " + 7	0		9		4		11		24	
+ 2 " + 4	2		9		8		7		26	
+ 1 " - 1	1	33	6	20	7	40	9	37	23	30
- 2 " - 4	1		5		5		7		18	
- 5 " - 7	1		12		6		3		22	
- 8 " -10	1		14		2		4		21	
-11 " -13	2		6		5		0		13	
-14 " -16	1	58	6	49	2	30		11	9	35
-17 " -19	1		3		0				4	
-20 " -22	0		4						4	
-23 " -25	1		2						3	
-26 " -28	0		1						1	
-29 " -30			1						1	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Die fehlenden Korrelationen zwischen den beiden Tests für die SS und HS können so gedeutet werden, daß zwischen den von den beiden Tests gemessenen Merkmalen resp. Fähigkeiten keine Gemeinsamkeit besteht. Wertet man die Korrelation als Maß der Merkmalsstabilität, so ist diese für diese beiden Gruppen gleich Null, jedoch immer unter der Voraussetzung, beide Tests messen das/die gleiche(n) Merkmal(e). Für die Verteilungen der SS und HS lassen sich somit auch keine spezifischen Erwar-

tungen formulieren. Um so bemerkenswerter ist der relativ hohe Anteil negativer Veränderungen sowohl bei den SS (58%) als auch bei den HS (49%) um mehr als -4 Punkte. Dieser Befund ist in der Tendenz dem unter e) vergleichbar.

- (2) Auch für die RS sind die relativen Veränderungen insgesamt mit denen unter e) vergleichbar.
- (3) Eine Abweichung findet sich bei den OS. 52% verbessern ihren späteren Testwert um mehr als +4 Punkte, während nur 11% eine negative Veränderung oberhalb -4 Punkte aufweisen.
- (4) Insgesamt finden sich hier - absolut gesehen- die größten positiven wie negativen Veränderungen.
- (5) Die Verteilungsprüfung für die HS, RS und OS (9 Felder) ist im Ergebnis hoch signifikant von Null verschieden. Bezieht man wegen fehlender Korrelationen bei den HS nur die RS und OS (6 Felder) in die Verteilungsprüfung ein, so ist die Gleichverteilungsabweichung zu Gunsten der OS auf dem 2%-Niveau bedeutsam.

h) Veränderung zwischen Intelligenztestleistungen und
Schulreifetestergebnissen: $GSW_{14} - KERNschulreife-$
testgesamtwert

Tabelle 28 zeigt die Verteilungen der Differenzen zwischen dem Gesamtstandardwert GSW 13-14 und dem Gesamtwert im KERNschulreifetest.

Tabelle 28: Häufigkeiten der Standardwertdifferenzen GSW¹⁴ -
KERNschulreifetestgesamtwert nach Schulart
getrennt

Differenz	SS		HS		RS		OS		Ges.	
	f	%	f	%	f	%	f	%	f	%
+32 bis +33			1		0		0		1	
+29 " +31			0		0		0		0	
+26 " +28			0		0		0		0	
+23 " +25			0		0		1		1	
+20 " +22			1		1		3		5	
+17 " +19			6		2		3		11	
+14 " +16		8	4	28	0	40	4	52	8	36
+11 " +13	1		5		3		6		15	
+ 8 " +10	0		7		3		6		16	
+ 5 " + 7	0		4		11		9		24	
+ 2 " + 4	0		12		5		7		24	
+ 1 " - 1	3	33	10	29	8	32	8	34	29	31
- 2 " - 4	1		7		3		6		17	
- 5 " - 7	0		9		5		5		19	
- 8 " -10	1		12		5		2		20	
-11 " -13	4		6		1		1		12	
-14 " -16	1		5		2		0		8	
-17 " -19	0	58	6	43	1	28	1	14	8	33
-20 " -22	0		2		0		0		2	
-23 " -25	1		2						3	
-26 " -28	0		0						0	
-29 " -30	0		1						1	
	12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) Wie schon unter g) ist die Anzahl negativer Veränderungen um mehr als -4 Punkte auch hier bei den SS (58%) und HS (43%) relativ am größten.
- (2) Ebenso sind die positiven Veränderungen um mehr als +4 Punkte bei den OS (52%) relativ am größten. Im genannten Bereich überwiegen auch bei den RS (40%) die Positiven Veränderungen.
- (3) Die Verteilungsprüfung (9 Felder) ist sehr signifikant. Bezieht man aus o. a. Grunde nur die RS und OS in diese Prüfung ein, so ergibt sich kein

signifikantes Ergebnis, wenngleich auch hier die positive Tendenz zugunsten der OS ausfällt.

i) Differenzen zwischen den Schulreifetestgesamtwerten:
MÜNCHNERgesamtwert - KERNgesamtwert-----

Schließlich sollen noch die Verteilungen der individuellen Differenzen der Schüler in den beiden Schulreifetests dargestellt werden. Es ist jedoch zu beachten, daß

- beide Tests jeweils am gleichen Tag durchgeführt wurden, was erforderlich machen könnte, etwaige Effekte anders zu deuten als sich aus den vorherigen Vergleichen ergibt;
- die Testerhebungen für die 1960 von der Einschulung zurückgestellten Schüler erst nach deren Einschulung im Jahre 1961 erfolgte. Dieser Tatbestand ist besonders für das/die vom KERN'schen Schulreifetest zu messende(n) Merkmal(e) von Bedeutung (vgl. KEMMLER 1967, S. 140f). Auch eine Beeinflussung der entsprechenden Korrelationen scheint wahrscheinlich.

Tabelle 29 zeigt die Häufigkeiten individueller Differenzen, die die vier Gruppen in den beiden Schulreifetestgesamtwerten aufweisen. Es wird die Differenz MÜNCHNERgesamtwert minus KERNgesamtwert gebildet.

Tabelle 29: Häufigkeiten der Standardwertdifferenzen
MÜNCHNERSchulreifetestgesamtwert - KERNSchul-
reifetestgesamtwert nach Schulart getrennt

Differenz		SS		HS		RS		OS		Ges.	
		f	%	f	%	f	%	f	%	f	%
+20				1		0		0		1	
+17	bis +19			1		1		2		4	
+14	" +16			5		2		1		8	
+11	" +13	1	17	4	23	3	22	8	48	16	30
+8	" +10	0		3		3		12		18	
+5	" +7	1		9		2		7		19	
+2	" +4	2		14		7		8		31	
+1	" -1	1	50	13	42	12	50	3	32	29	40
-2	" -4	3		15		6		9		33	
-5	" -7	3		12		6		8		29	
-8	" -10	1		12		3		1		17	
-11	" -13		33	3	35	3	28	2	19	8	29
-14	" -16			4		2		1		7	
-17	" -19			3		0		0		3	
-20				1		0		0		1	
		12	100	100	100	50	100	62	100	224	100

Ergebnisse:

- (1) SS, HS und RS weisen die relativ größten Übereinstimmungen in den beiden Tests auf, d. h. bei 42 - 50% der Schüler dieser Gruppen beträgt die Abweichung von Test zu Test maximal ± 4 Punkte.
- (2) Für die gleichen Gruppen Überwiegen ferner die relativen Anteile negativer Differenzen (mehr als -4 Punkte) gegenüber den positiven im analogen Bereich.
- (3) Von diesen in den Gruppen der SS, HS und RS tendenziell gleichen Verteilungen weichen die OS stark ab. Während 48% OS positive Differenzen um mehr als +4 Punkte erzielen, liegt der relative Anteil negativer Differenzen bei nur 19%. Ca.

2 Drittel der OS weisen somit Differenzen in den Extrembereichen auf, gegenüber maximal 50% in den übrigen Gruppen.

- (3) Es fällt auf, daß bei mehr als $1/5$ der Probanden Testdifferenzen von mehr als $\pm$ einer bis zu $\pm$ zwei Standardwerteinheiten vorkommen; 13% der Probanden erzielen im KERNSchulreifetest ein um mehr als eine Standardwerteinheit höheres, 8 % ein um mehr als eine Standardwerteinheit niedrigeres Resultat.
- (4) Die Gleichverteilungsprüfung (9 Felder, ohne SS) ist auf dem 1%-Niveau signifikant, d. h. das insgesamt positivere Abschneiden der OS bzw. negativere der HS muß auf (noch) zu identifizierende Faktoren zurückgeführt werden.

8.2

Veränderungen nach Schülergruppen zusammengefaßt

Allgemein ist festzustellen, daß Veränderungen der verschiedenen Testleistungen (Gesamt- und spezifische Begabung) interschulisch differieren, während intraschulisch eine gewisse Systematik vorliegt.

Sonderschüler:

Die Leistungsänderungen der SS in allen Tests bewegen sich größtenteils im mittleren und negativen Bereich, d.h. die Tatsache des Sonderschulbesuchs - mit spezifischer individueller Förderung (?) - wirkt sich auch auf spezifische Begabungen (soweit sie von den verwendeten Tests erfaßt werden) für die SS nicht positiv aus! Für mögliche Gründe s.u..

Hauptschüler:

Die HS zeigen ein differenzierteres Bild. Unter ihnen finden sich Schüler, die die absolut größten positiven Veränderungen im Gesamtstandardwert, USW-Verbal, USW-Reasoning und USW-Lesefrei aufweisen. Im USW-Verbal und im USW-Reasoning stellt diese Gruppe allerdings auch die Schüler mit den absolut größten negativen Veränderungen. Bezogen auf die extremen Veränderungen ist bei den HS festzustellen, daß im USW-Reasoning mit 39% die nega-

tiven Veränderungen überwiegen, gegenüber 23% positiver Veränderungen. Umgekehrt sind im USW-Lesefrei mit 36% die meisten positiven Veränderungen zu finden, gegenüber nur 17% negativer Veränderungen. Dagegen weichen die relativen extremen positiven wie negativen Veränderungen im Gesamtstandardwert und USW-Verbal nur um 1% voneinander ab bzw. sind genau gleich groß.

Beim Vergleich der GSW 9-10 bzw. GSW 13-14 mit den Schulreifestesgesamtwerten ergibt sich folgendes: Sowohl beim MÜNCHNER- als auch beim KERNschulreife-test gibt es in allen Fällen in den Extrembereichen die relativ größten Anteile bei den negativen Veränderungen. Sie schwanken zwischen 8% (GSW 13-14/MÜNCHNER) und 18% (GSW 9-10/KERN). Diese negativen Veränderungen finden sich auch in den Extrembereichen der Differenzen zwischen den beiden Schulreifestests. 35% der Schüler weisen in den Extrembereichen der Schulreifestesleistungen im MÜNCHNER-Gesamtwert eine negative und nur 23% eine positive Differenz auf.

Realschüler:

Bei den RS überwiegen - bei einzelnen Tests im Gegensatz zu den HS - in den Extrembereichen der Tests, die besonders das schulische Fortkommen

messen sollen, die positiven relativen Frequenzen; so im Gesamtstandardwert und USW-Verbal mit 28% bzw. 30%, bei nur jeweils 14% negativer Veränderungen. Im USW-Reasoning zeigen mit 32% noch 8% mehr Schüler eine positive Veränderung. Im USW-Lesefrei sind die relativen Häufigkeiten in den Extrembereichen quasi ausgeglichen. Ein Blick auf die absolut größten Veränderungen zeigt: Unter den RS ist einer der Schüler mit der insgesamt größten negativen Veränderung. Im USW-Verbal und USW-Lesefrei erzielt je ein Schüler die größten positiven Veränderungen. Im USW-Reasoning liegen die RS mit den größten Veränderungen unterhalb der Absolutheitsgrenze für die Gesamtstichprobe.

Oberschüler:

Im Mittelbereich (± 4 Punkte) weichen die OS kaum von der Gesamtpopulation ab. Beträchtliche, entgegen der Regression zur Mitte verlaufende Tendenzen finden sich im Bereich des Verbalen, Reasoning, Lesefrei und den Schulreife-tests mit den Gesamtstandardwerten: So verbessern sich 5% mehr Schüler als zu erwarten im Verbalen, im Reasoning sind es sogar 22% mehr. Dagegen verschlechtern sich im Skore Lesefrei 29% mehr Schüler als zu erwarten gewesen wäre. In allen Vergleichen der Schulreife-tests mit den GSW's sind die positiven Veränderungen größer als die negativen. Die positiven

Differenzen schwanken zwischen 8% und 41%, wobei die Unterschiede des GSW's mit dem MÜNCHNER geringer ausfallen als mit dem KERN. Das deutet u.E. darauf hin, daß der MÜNCHNER eher einem Intelligenztest nahekommmt als der KERN (vgl. auch KEMMLER 1967, 1976).

Insgesamt deuten unsere Ergebnisse darauf hin, daß das Phänomen der Regression zur Mitte einer differenzierteren Analyse durchaus zugänglich ist.

8.3 Mittelwertdifferenzen und Regression zur Mitte

Bei der Diskussion um die Intelligenz wird der IQ als eine besonders relevante Größe betrachtet, so daß auch wir an seiner Verwendung nicht immer vorbeikommen. Dennoch geben wir jedoch der bereits 1911-13 von ROSSOLIMO vorgeschlagenen Methode der Profildarstellung kognitiver Leistungen den Vorzug, da auf diese Weise individuelle Stärken und Schwächen der Probanden genauer zum Ausdruck kommen können, wie wir hoffen, mit Hilfe von Diskriminanzanalysen weiter unten demonstrieren zu können. Da es sich beim IQ, gemessen in Testleistungen, um eine multifaktorielle, also zusammengesetzte Größe handelt, erhebt sich die Frage, ob die Regression zur Mitte bei genau umschriebenen Untergruppen eines Altersjahrgangs immer in gleicher Weise auftritt, oder ob sich mit Blick auf spezifische "Primärfaktoren" der Intelligenz (hier: sensu THURSTONE) Abweichungen von der statistisch zu erwartenden Tendenz zur Mitte ergeben und wie diese erklärt werden können.

Eine Vorbedingung für dieses Vorgehen, nämlich ausreichend hohe Reliabilitäten der Tests, ist u.E. zufriedenstellend erfüllt, bewegen sich doch die Koeffizienten zwischen 0,76 und 0,98.

Wir haben nun für die Standardwerte (ganzer Jahrgang) der Schulreifetests 1960/61, der Unterstandardwerte (USW) und der Gesamtstandardwerte (GSW) der Untersuchungen 1964 und 1968 Veränderungswerte berechnet (WILCOXON-Rangtest für korrelierte Stichproben), deren Hauptergebnisse in den folgenden Tabellen dargestellt und erläutert werden.

Berechnet wurden die Werte für folgende Gruppen:

- a) Ältere/Jüngere,
- b) Jungen/Mädchen,
- c) Altergemäß Eingeschulte/Zurückgestellte,
- d) Altergemäß eingeschulte Jungen/Mädchen,
- e) Zurückgestellte Jungen/Zurückgestellte Mädchen,
- f) Altergemäß eingeschulte ältere Jungen/Mädchen,
- g) Altergemäß eingeschulte jüngere Jungen/Mädchen,
- h) SS+HS/RS+OS,
- i) SS/HS,
- j) RS/OS,
- k) OS mit $GSW_{64} > 67$ /HS mit $GSW_{64} < 37$ Standardwertpunkte.

Abgesehen von den Vergleichen der beiden Schulreifetests mit den GSW_{64} und GSW_{68} sind bei den Gruppen a) bis g) und k) die Regressionen zur Mitte wie statistisch zu erwarten. Die Schulreifetests sind

zwar in den Tabellen enthalten, sollen hier nicht besonders interpretiert werden, weil die Reliabilitätswerte überhaupt oder in ihrem Zustandekommen nicht hinreichend bekannt sind. Es folgen nun die Darstellungen nach der besuchten Schule, also SS+HS, RS+OS, sowie getrennt SS, HS RS und OS.

1) SS+HS und RS+OS

Sowohl bei den SS+HS als auch bei den RS+OS gibt es im Verbal-Faktor und bei den GSW's keine Veränderungen zwischen den Untersuchungen 1964 und 1968. Anders beim Reasoning-Faktor: Hier gibt es bei beiden Gruppen erwartungswidrige Signifikanzen (5%). Die bei der Untersuchung 1964 unter dem Durchschnitt liegenden SS+HS sollten gemäß der statistischen Regressionserwartung 1968 einen höheren, näher dem Mittelwert der Gesamtpopulation liegenden Wert erzielen. Umgekehrt bei den RS+OS; sie hätten 1968 einen geringeren mittleren Reasoning-Faktor erreichen sollen. Tatsächlich hat hier jedoch ein Zuwachs stattgefunden. Die auf dem 5%-Niveau signifikanten Differenzen widersprechen in ihrer Höhe damit genau dem Regression-zur-Mitte-Phänomen, d.h. die RS+OS haben 1968 hinzugewonnen, die SS+HS dagegen abgenommen.

Tabelle 30 : Differenzen "analoger" T-Standardwerte (WILCOXON-Test) für korrelierte Stichproben)

Normmittelwert = 50

Testdifferenzen zwischen ...	SS+HS			RS+DS		
	N = 112		p%	N = 112		p%
	AM	AM		AM	AM	
KERN-MÜNCHNER	46,8	45,7		52,6	54,3	5
GSW64-KERN	44,2	46,8	5	55,9	52,6	1
GSW68-KERN	44,1	46,8	5	55,9	52,6	1
GSW64-MÜNCHNER	44,2	45,7		55,9	54,3	
GSW68-MÜNCHNER	44,1	45,7		55,9	54,3	
VERBAL68-VERBAL64	43,4	44,0		56,6	56,0	
REASONING68-REASONING64	44,0	45,6	5	55,9	54,5	5
LESEFREI68-LESEFREI64	46,4	44,8	5	53,6	55,3	5
GSW68-GSW64	44,1	44,2		55,9	55,9	
VERBAL64-LESEFREI68	44,0	46,4	5	53,6	56,0	5
REASONING64-GSW64	45,6	44,2	1	54,5	55,9	5
REASONING64-VERBAL68	45,6	43,4	1	54,5	56,6	5
LESEFREI64-GSW64	44,8	44,2	5	55,3	55,9	5
LESEFREI68-GSW64	46,4	44,2	1	53,6	55,9	1
LESEFREI68-VERBAL68	46,4	43,4	1	53,6	56,6	1
LESEFREI68-REASONING68	46,4	44,0	1	53,6	55,9	5
LESEFREI68-GSW68	46,4	44,1	1	53,6	55,9	1

p% entspricht der 5% bzw. 1% Irrtumswahrscheinlichkeit

AM = Arithmetischer Mittelwert

Wie kann man sich nun das Zustandekommen besonders der erwartungswidrigen Ergebnisse erklären? Wir meinen, daß dies vermutlich am ehesten mit den spezifischen Primäranforderungen der verschiedenen Schularten zusammenhängt, d.h. während Sonder- und Hauptschüler hauptsächlich mit praktischen Aufgaben konfrontiert werden, werden Realschüler und noch mehr die Oberschüler (Gymnasiasten) während der gesamten Schulzeit mit immer neuen Problemen, die schlußfolgerndes Denken bzw. rationales Handeln erfordern, konfrontiert. Dies wird noch mal besonders unterstrichen, wenn man für beide Gruppen die Differenzen der "Faktoren" Lesefrei 68 - Reasoning 68 anschaut. Während die SS+HS sich sehr signifikant im sprachfreien Bereich verbessern, erzielen die RS+OS einen erhöhten Reasoningwert. Auch dieses Ergebnis scheint uns nur schwer in Einklang mit der Regression zur Mitte zu bringen zu sein, wenn die Ursache hierfür nicht in der Testkonstruktion selbst zu suchen ist.

2) Sonderschüler

Es kann festgestellt werden, daß die oben beschriebenen Effekte nicht von den SS verursacht werden, obwohl sie sich doch am unteren Ende der Skalen befinden und getreu der Regression zur Mitte am ehesten

Tabelle 31 : Differenzen "analoger" T-Standardwerte (WILCOXON-Test) für korrelierte Stichproben)

Normmittelwert = 50

Testdifferenzen zwischen ...	SS			HS		
	N = 12		p%	N = 100		p%
	AM	AM		AM	AM	
KERN-MÜNCHNER	37,5	36,7		47,9	46,8	
GSW64-KERN	31,2	37,5		45,7	47,9	
GSW68-KERN	30,3	37,5	5	45,8	47,9	
GSW64-MÜNCHNER	31,2	36,7		45,7	46,8	
GSW68-MÜNCHNER	30,3	36,7	1	45,8	46,8	
VERBAL68-VERBAL64	31,7	34,0		44,8	45,2	
REASONING68-REASONING64	33,3	33,3		45,3	47,1	5
LESEFREI68-LESEFREI64	31,5	31,4		48,2	46,4	5
GSW68-GSW64	30,3	31,2		45,8	45,7	
VERBAL64-GSW64	34,0	31,2	5	45,2	45,7	
VERBAL64-LESEFREI68	34,0	31,5	5	45,2	48,2	
REASONING64-GSW64	33,3	31,2		47,1	45,7	1
REASONING64-VERBAL68	33,3	31,7		47,1	44,8	1
LESEFREI64-GSW64	31,4	31,2		46,4	45,7	5
LESEFREI68-GSW64	31,5	31,2		48,2	45,7	1
LESEFREI68-VERBAL68	31,5	31,7		48,2	44,8	1
LESEFREI68-REASONING68	31,5	33,3		48,2	45,3	1
LESEFREI68-GSW68	31,5	30,3		48,2	45,8	1

p% entspricht der 5% bzw. 1% Irrtumswahrscheinlichkeit

AM = Arithmetischer Mittelwert

hier hätten die Werte positiv beeinflussen sollen. Es scheint so zu sein, wie schon von anderen Autoren festgestellt, daß sich die SS in ihrer kognitiven Entwicklung mit dem Übergang zur Sonderschule nicht mehr positiv verändern, und die Sonderschulen nur der Aufnahme von "abgeschobenen" Schülern dient und nicht etwa der individuellen kognitiven Förderung, die später wieder einen Vergleich mit Schülern anderer Schulen zuließe. Bemerkenswert erscheint sogar die Abnahme im Verbal-Faktor, die nur wegen der geringen Probandenzahl hier nicht signifikant ist.

3) Hauptschüler

Die Werte in Tabelle 31 zeigen deutlich, daß der oben beschriebene Effekt nur von den Hauptschülern hervorgerufen wird. Der besondere Zuwachs im "Faktor" Lesefrei 68 wird noch weiter deutlich, wenn man ihn vergleicht mit Verbal 68, Reasoning 68 und GSW 68. In jedem Falle dominiert Lesefrei 68 sehr signifikant über die drei übrigen Werte, auch über Verbal 64 und GSW 64.

4) Realschüler

Wie die Werte in Tabelle 32 zeigen, haben sich die RS nicht verändert. Vom Phänomen der Regression

zur Mitte waren auch hier die geringsten Veränderungen zu erwarten, da diese Schüler lediglich knapp überdurchschnittliche Mittelwerte erzielt haben. Immerhin sollte nicht unerwähnt bleiben, daß die RS sich im Reasoning-Faktor um 1,2 SW-Punkte und im GSW um 0,7 SW-Punkte verbessert haben.

5) Oberschüler (Gymnasiasten)

Da die RS zu dem Effekt, den wir bei den RS+OS gefunden haben, nichts beigetragen haben, muß er folglich allein von den OS bewirkt werden. Dies würde dann unsere Hypothese stützen, wonach schlußfolgerndes Denken gleichsam der rote Faden ist, der sich als Ergebnis durch alle Unterrichtsfächer der Gymnasialschüler zieht. Tabelle 32 zeigt, daß auch hier die Leistungen im Verbal-Faktor und in den GSW's über vier Jahre, resp. im Alter von 10 und 14 Jahren konstant sind, dagegen im Reasoning-Faktor ein erwartungswidriges und im "Faktor" Lesefrei ein erwartungsgemäßes Resultat zeigen. Die Oberschüler, obgleich schon 1964 überdurchschnittlich, werden im Reasoning-Faktor noch besser, im "Faktor" Lesefrei tendieren sie dagegen zur Populationsmitte. Es sind also im wesentlichen die schon beim Übergang zur Oberschule vorhandene Verbalität und die sich noch steigernde Denkfähigkeit, die den Erfolg der Gymnasiasten bestimmen.

Tabelle 32 : Differenzen "analoger" T-Standard-
werte (WILCOXON-Test für korre-
lierte Stichproben)
Normmittelwert = 50

Testdifferenzen zwischen ...	RS			OS		
	N = 50		p%	N = 62		p%
	AM	AM		AM	AM	
KERN-MÜNCHNER	51,9	51,7		53,2	56,3	1
GSW64-KERN	52,3	51,9		58,7	53,2	1
GSW68-KERN	53,0	51,9		58,3	53,2	1
GSW64-MÜNCHNER	52,3	51,7		58,7	56,3	
GSW68-MÜNCHNER	53,0	51,7		58,3	56,3	
VERBAL68-VERBAL64	53,7	52,5		58,9	58,8	
REASONING68-REASONING64	52,5	52,1		58,6	56,4	5
LESEFREI68-LESEFREI64	52,0	52,2		55,0	57,8	1
GSW68-GSW64	53,0	52,3		58,3	58,7	
VERBAL64-LESEFREI68	52,5	52,0		58,8	55,0	1
REASONING64-GSW64	52,1	52,3		56,4	58,7	5
REASONING64-GSW68	52,1	53,0		56,4	58,3	5
LESEFREI64-GSW64	52,2	52,3		57,8	58,7	5
LESEFREI68-GSW64	52,0	52,3		55,0	58,7	1
LESEFREI68-VERBAL68	52,0	53,7		55,0	58,9	5
LESEFREI68-REASONING68	52,0	52,5		55,0	58,6	1
LESEFREI68-GSW68	52,0	53,0		55,0	58,3	1

p% entspricht der 5% bzw. 1% Irrtumswahrscheinlichkeit

AM = Arithmetischer Mittelwert

Abschließend wollen wir noch einige Daten wiedergeben, die wir nach einer Arbeit von FIEGENBAUM (1974) zusammengestellt haben. Er hat alle erreichbaren Oberschüler bis zum Ausscheiden aus der Schule verfolgt. Die nachfolgende Tabelle 33 enthält die Schüler, die im Sommer 1972 nochmals den Intelligenz-Struktur-Test (IST 70) bearbeitet haben, so daß von dieser Gruppe drei Gesamtstandardwerte miteinander verglichen werden können. Wie man sieht, können gut die Mittelwertsveränderungen als Regression zur Mitte gedeutet werden. Eine Ausnahme bilden die Schüler der Klasse 12. Ihr gegenläufiger Effekt ist dadurch verständlich, daß sie in den PMA die niedrigsten, jeweils nahe am Durchschnitt liegende Werte erzielt haben, so daß eine Steigerung nicht ganz unwahrscheinlich sein muß.

Tabelle 33 : Regressionseffekte bei den Abiturienten im Gesamtstandardwert 1964, im Gesamtstandardwert 1968, im Intelligenz-Struktur-Test IST 70

Status am 1. 8. 1972	N	GSW64		GSW68		IST 70	
		AM	St	AM	St	AM	St
Abitur	21	64,7	5,3	63,4	6,8	60,2	4,9
Klasse 13	10	58,1	5,6	56,6	7,0	53,9	7,7
Klasse 12	7	53,7	5,8	55,3	5,4	56,6	9,1
Klasse 11	2	60,0	7,1	57,0	12,7	51,0	2,8
	40	60,9	6,9	60,0	7,6	57,5	7,0

Die beiden Schüler der Klasse 11 und ein Schüler der Klasse 13 haben das Abitur nicht erreicht.

8.4 Extreme Aufsteiger und Absteiger im GSW

In unserer Stichprobe gibt es ca. 7% Schüler (= 16), die sich im Gesamtstandardwert in vier Jahren im Alter von 10 - 14 Jahren signifikant verändert haben. Die signifikante Veränderung definieren wir wie folgt:

Zugrundegelegt wurde die Differenz zwischen GSW 13-14 minus GSW 9-10. Als Signifikanztest benutzen wir das Konfidenzintervall zweier Testwerte (vgl. LIENERT² 1967, S. 458f)

$$U_{\text{crit}} = z_{\text{crit}} \cdot s \left(\sqrt{1 - r_{11}} + \sqrt{1 - r_{22}} \right)$$

z_{crit} = 1,96 für Irrtumswahrscheinlichkeit $p < 5\%$

s = 10 Standardabweichung

r_{11} = 0,94 Reliabilität des GSW 9-10

r_{22} = 0,95 Reliabilität des GSW 13-14

$$U_{\text{crit}} = 9,2 \text{ GSW-Punkte}$$

Die Durchsicht unserer Daten ergab nun, daß sieben Probanden eine positiv-signifikante Veränderung von mehr als +9 GSW-Punkten und neun eine negativ-signifikante Veränderung von mehr als -9 GSW-Punkten aufweisen. Im folgenden soll nun versucht werden herauszufinden, ob sich bei den Probanden bestimmte Merkmale kumulieren, so daß man von homogenen Gruppen

oder Subgruppen sprechen kann, bzw. ob Merkmale außerhalb des Regression-zur-Mitte Effekts (s. u.) gefunden werden können, die für die Veränderung (mit-)ursächlich gewesen sind.

In Tabelle 34 haben wir einige relevante Merkmale der Aufsteiger und Absteiger (bei MANDL 1975 übrigens Faller, Bleiber und Steiger genannt; zuweilen werden in der deutschen Literatur auch die englischen Begriffe "Gainer" und "Loser" verwendet z.B. PETERMANN 1978, S. 29) zusammengestellt. Daß bei den Aufsteigern die HS und bei den Absteigern die OS dominieren, weist deutlich auf das Phänomen der Regression zur Mitte hin. Unterschiede im Geschlecht, Alter, Beruf des Vaters und Einschulung liegen genau im Bereich des Zufalls.

Auf welche Leistungen in welchen Tests sind nun die Unterschiede der beiden Gruppen zurückzuführen? Die Tabellen 35, 36 und 37 enthalten Mittelwerte, Streuungen, Minimal- und Maximaltestwerte sowie die Irrtumswahrscheinlichkeiten von Gruppenunterschieden. Wie zu erwarten, kommen die signifikanten Unterschiede infolge der starken Abweichungen in den Tests 1964 zustande, was noch besonders dadurch unterstrichen wird, daß die Mittelwerte 1968 völlig gleich sind. Die Abbildung 1 soll diese etwas unanschaulichen Ergebnisse deutlicher werden lassen. Die Regression zur Mitte ist dabei unverkennbar.

Tabelle 34 : Aufsteiger(+) und Absteiger(-) mit GSW68 - GSW64 > 191

lfd.-Nr.	Schul- art	GSW verän- derung	Schul- ab- schluß	Ge- schlecht	Alters- gruppe	Berufsgruppe des Vaters	Ein- schu- lung	gew. Noten D M G
302	OS	+12	DO	m	jünger	4, manuelle Berufe	'61	7 6 6
21	RS	+10	MR	w	Älter	5, Ungelernte	'60	4 4 4
37	HS	+11	HA	w	Älter	4, manuelle Berufe	'60	4 4 4
220	HS	+10	HA	w	Älter	3, "Stehkragenberufe"	'60	5 6 5
230	HS	+12	HA	m	jünger	4, manuelle Berufe	'60	5 6 5
313	HS	+14	HA?	w	jünger	4, manuelle Berufe	'61	8 7 8
329	HS	+14	HA?	m	jünger	2, mittl.Führungspos.	'61	5 7 6
109	OS	-11	Abi	m	Älter	2, mittl.Führungspos.	'60	4 4 4
145	OS	-15	Abi	m	jünger	4, manuelle Berufe	'60	2 4 3
177	OS	-11	Abi	m	Älter	3, "Stehkragenberufe"	'60	3 3 3
193	OS	-12	Abi	m	jünger	1, Hochschulabschluß	'60	3 3 3
144	RS	-10	MR	m	jünger	3, "Stehkragenberufe"	'60	6 7 6
180	RS	-14	?	w	jünger	2, mittl.Führungspos.	'60	6 6 6
46	HS	-13	HA	w	jünger	3, "Stehkragenberufe"	'60	4 6 5
168	HS	-12	HA	w	Älter	4, manuelle Berufe	'60	4 5 5
253	HS	-11	HA?	m	jünger	3, "Stehkragenberufe"	'60	4 6 5

DO = Drop out; HA = Hauptschulabschluß; MR = Mittlere Reife; Abi = Abitur
 Noten: O = Deutsch; M = Mathematik; G = Mittel aus D und M

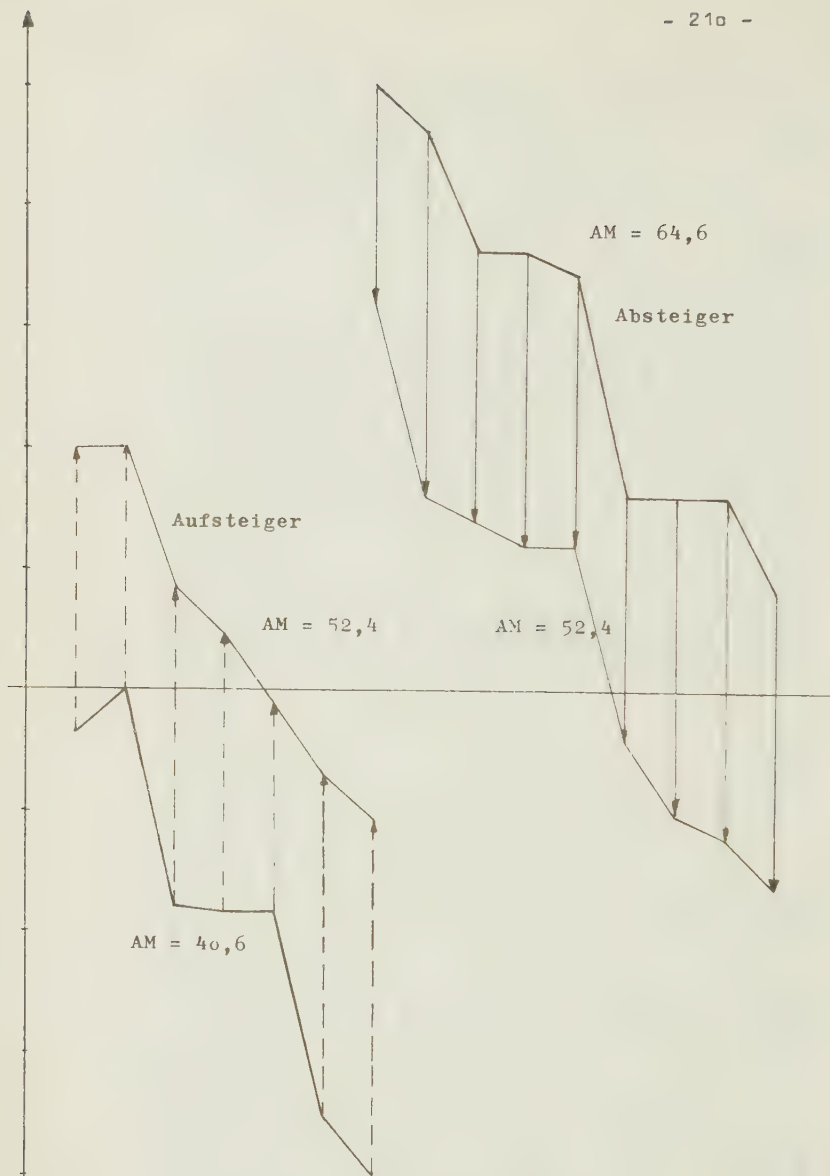


Abb. 1: Aufsteiger und Absteiger im

GSW 13-14 - GSW 9-10

(Normittelwert = 50)

Tabelle 35 : Standardwerte der Gesamtskores, USW
und GSW der Tests 1960/61, 1964
und 1968

Normmittelwert = 50

Test		AM	St	Min	Max	P%
KERN-Ges.	1	51,8	7,5	43	61	
	2	51,7	9,4	37	61	
MÜNCHNER-Ges.	1	53,3	8,9	39	62	
	2	47,7	6,9	41	61	
VERBAL64	1	60,6	10,5	47	77	5
	2	37,3	9,4	23	49	
REASONING64	1	59,7	7,4	49	73	5
	2	44,0	7,3	33	53	
LESEFREI64	1	64,4	7,6	54	77	0,1
	2	41,1	6,6	33	52	
GSW 64	1	64,6	7,9	54	77	0,1
	2	40,6	7,0	31	50	
VERBAL68	1	56,7	11,1	38	72	5
	2	43,1	8,1	29	52	
REASONING68	1	50,4	10,3	33	65	
	2	50,1	6,2	44	59	
LESEFREI68	1	52,0	6,7	39	63	
	2	56,4	6,2	47	67	
GSW 68	1	52,4	8,0	42	66	
	2	52,4	6,0	45	66	

Gruppe 1 := Absteiger (Differenz - 9; N = 9)

Gruppe 2 := Aufsteiger (Differenz + 9; N = 7)

AM := Arithmetisches Mittel

St := Standardabweichung

Min. := Minimalwert

Max. := Maximalwert

Tabelle 36 : Standardwerte des PMA 9-10 und vier weiterer Leistungstests

Normmittelwert = 50

Test		AM	ST	Min	Max	P%
BILDERWORTSCHATZ	1	60,0	11,5	42	77	5
	2	35,7	9,3	23	48	
WORTSCHATZ	1	58,1	8,1	47	70	5
	2	41,9	10,2	26	53	
RAUMERFASSUNG	1	61,1	11,3	49	77	5
	2	49,0	3,4	45	54	
FIGURENGRUPPEN	1	58,2	6,5	51	71	5
	2	45,7	6,7	37	55	
WORTGRUPPEN	1	57,9	8,7	46	74	5
	2	43,9	10,7	28	59	
WAHRNEHMUNGSSCHN.	1	62,6	9,5	45	77	5
	2	42,9	12,7	23	57	
ZAHLEN	1	59,1	7,4	46	67	5
	2	45,3	9,5	34	57	
WISSENSTEST	1	59,0	6,5	50	66	
	2	48,3	10,2	33	57	
FEHLERSUCHEN	1	54,0	3,2	49	58	
	2	50,3	9,4	32	61	
WÖRTERSUCHEN	1	56,2	9,1	46	67	
	2	48,6	9,3	34	63	
WÖRTERTRENNEN	1	58,8	11,6	40	73	
	2	47,9	8,8	33	62	

Gruppe 1 := Absteiger (Differenz -9; N = 9)

Gruppe 2 := Aufsteiger (Differenz +9; N = 7)

AM := Arithmetisches Mittel

St := Standardabweichung

Min. := Minimalwert

Max. := Maximalwert

Tabelle 37 : Standardwerte der PMA 13-14 und fünf weiterer Leistungstests

Normmittelwert = 50

Test		AM	ST	Min	Max	P%
WORTSCHATZ	1	56,7	11,1	38	72	5
	2	43,1	8,1	29	52	
FIGURENVERGLEICH	1	48,9	5,7	39	57	
	2	54,6	4,0	48	60	
BUCHSTABENREIHEN	1	48,6	8,6	32	58	
	2	50,9	9,4	42	66	
WORTGRUPPEN	1	52,7	10,1	40	68	
	2	49,7	3,9	43	55	
WAHRNEHMUNGSSCHN.	1	54,4	10,0	35	65	
	2	50,7	7,3	40	59	
ZAHLEN	1	50,9	7,4	39	58	5
	2	60,3	6,5	52	68	
WORTANALOGIEN	1	52,9	8,7	40	67	5
	2	45,0	6,7	39	55	
TEXTAUFGABEN	1	50,9	6,3	41	60	
	2	45,9	4,4	39	50	
FIGURENGRUPPEN	1	50,3	8,5	40	64	
	2	47,4	8,2	40	59	
KLT 30	1	53,1	6,3	40	62	
	2	57,7	6,0	49	66	
DIKTAT	1	56,6	7,7	46	71	
	2	49,3	9,3	32	61	

Gruppe 1 := Absteiger (Differenz -9; N = 9)

Gruppe 2 := Aufsteiger (Differenz +9; N = 7)

AM := Arithmetisches Mittel

St := Standardabweichung

Min. := Minimalwert

Max. := Maximalwert

Bei den einzelnen Leistungstests fällt auf, daß sich die Aufsteiger und Absteiger in allen zu den PMA 9-10 gehörenden Tests signifikant unterscheiden, nicht dagegen in den Tests Wissenstest, Fehlersuchen, Wörtersuchen und Wörtertrennen.

Bei der Untersuchung 1968 erzielten die Absteiger in den Tests Wortschatz und Wortanalogien signifikant niedrigere Mittelwerte als die Aufsteiger. Tendenziell haben die Absteiger auch noch in den PMA-Tests Wortgruppen und Wahrnehmungsschnelligkeit höhere Mittelwerte behaupten können, von den übrigen Tests sind es Textaufgaben, Figurengruppen und Diktat.

Wie bereits erwähnt, waren wir uns wegen der geringen N's der Aufsteiger und Absteiger bewußt, daß es schwierig sein würde, signifikante Gruppenunterschiede in den durchgeführten Tests zu finden.

Schließlich wollten wir wissen, ob sich die Schüler in den Fragebogen-Antworten unterscheiden, und zwar:

- Skores des Kinderfragebogens 1964,
- ENRA-Fragebogen 1968,
- IAR-Fragebogen 1968,
- Interessen-Milieu-Fragebogen (IMF) 1968,
- Eltern-Milieu-Fragebogen (EMF) 1968,
- Lehrerfragebogen 1968.

Dazu haben wir einfache Varianzanalysen oder Chi-Quadrate berechnet: Von den sieben Skores des Kinderfragebogens (vgl. NIEMEIER 1968; KEMMLER 1976) sind drei interessant. Die Absteiger haben weniger "Angst" ($p < 1\%$) als die Aufsteiger. Sie beurteilen ihre "Schulleistungen" auch tendenziell ($p < 10\%$) schlechter und schätzen ihre "Schulanpassung" geringer ein als die Aufsteiger.

Im ENRA-Fragebogen (Extraversion, Neurotizismus, Rigidität, Aufrichtigkeit) unterscheiden sich beide Gruppen nicht.

Dagegen zeigen im Fragebogen für internale/externale Kontrolle (IAR von CRANDELL 1965), - Attribution von Erfolg bzw. Mißerfolg - die Absteiger sowohl im Gesamtwert (I_g) als auch im Mißerfolgswert (I_m) jeweils tendenziell ($p < 10\%$) niedrigere Werte als die Aufsteiger. Die Absteiger suchen die Ursachen

für Mißerfolg eher bei sich selbst als die Aufsteiger.

Sowohl im Interessen-Milieu-Fragebogen als auch im Elternfragebogen zeigen beide Gruppen keine Unterschiede oder der Tendenz nach Auffälligkeiten, die im Sinne spezifischer Probanden-"Pattern" zu interpretieren wären.

Selbst wenn man die Differenzgrenze von absolut 10 auf 7 Standardwertpunkte verändert, um eine größere Probandenzahl zu erhalten (hier: $N = 35$), führt dies nicht zu anderen Ergebnissen. Lediglich was die Zahl der Bücher betrifft (IMF), so haben die Absteiger signifikant mehr Bücher in ihrem Besitz als die Aufsteiger.

Zwei weitere, das Schulumfeld tangierende Aspekte ergeben sich im Lehrerfragebogen: Die Fragen 4 und 5 "Hat das Kind Kontakt zu seinen Klassenkameraden (oder Kameradinnen)?" und "Befolgt das Kind die Vorschläge und Anordnungen des Lehrers?" werden von den Lehrern so beantwortet (signifikant), daß die Absteiger weniger Kontakt zur Schulkameraden haben und sich häufig den Anordnungen des Lehrers widersetzen.

Unsere Probandengruppe ist viel zu klein, zudem gibt es auch methodologische Argumente, die uns hindern, den Gesamtbefund zu deuten. Wir wollen hier daher mit aller gebotenen Vorsicht nur vermuten, daß unsere Befunde geeignet sein könnten, bei Berücksichtigung des schulischen Umfeldes der Schüler in einer neuen Untersuchung des Phänomen der Regression zur Mitte so aufzuspalten, daß neben dem rein statistischen Aspekt auch der psychologisch-relevantere deutlicher ins Blickfeld kommt.

Schließlich sei noch hinzugefügt, daß in den analogen Verbal-Faktoren vier Probanden absolute Differenzen von 16-19 USW-Punkten erzielten. In den Reasoning-Faktoren betragen die Differenzen bei insgesamt sechzehn Personen 14-23 USW-Punkte. Und schließlich veränderten sich neunundvierzig Schüler zwischen 10-23 USW-Punkte in den Lesefrei-"Faktoren".

9.0 DIE DISKRIMINANZANALYSE ALS
DIAGNOSE- UND PROGNOSEHILFE

9.1 Vorbemerkung:

Ab Mitte der 60er Jahre wurde mit der verstärkten Bildungswerbung auch die Frage immer dringlicher, wie man denn die richtige, erfolgversprechende Schullaufbahn für die Schüler "bestimmen" könne. Die üblichen Verfahren erschienen zu subjektiv und waren häufig den Anforderungen nicht gewachsen. Der Ruf nach objektiven Tests wurde laut. Als man merkte, daß auch Tests ihre Grenzen in ihrer Validität haben, und gute Tests viel Geld kosten, also kaum zu haben sind, schlug das Pendel zur Abkehr von den "seelenlosen" Tests. Immerhin, so könnte man spötn, werden die schriftlichen Arbeiten in den Schulen heute nicht mehr Klassenarbeiten, sondern Tests genannt, die Objektivität sei also gewahrt.

De facto sind die neuen Probleme aber immer noch die alten. HUTHER schrieb 1910:

"Es ist bisher nicht gelungen, eine allgemein befriedigende Theorie der intellektuellen Begabung zu schaffen" (HUTHER 1910, S. 397).

Abgesehen davon, daß dieser Satz 70 Jahre später kein Iota von seiner Richtigkeit eingebüßt hat, hat er

auch leicht abgewandelt seine Richtigkeit:

Es ist bisher noch nicht gelungen, ein allgemein befriedigendes Inventarium zur Schullaufbahnberatung zu schaffen. Dieses Problem ist ebenso alt, wie das vorerwähnte. Daraus ergibt sich eigentlich schon, daß auch wir keine allseits befriedigende Lösung anzubieten haben. Was wir versuchen wollen ist, zu untersuchen, ob man der Lösung mit Hilfe eines in der (pädagogischen) Psychologie bisher nur selten eingesetzten Verfahrens, der Diskriminanzanalyse, ein wenig näher kommen kann.

Es sei schon hier angemerkt, daß dieser mathematisch-statistische Algorithmus natürlich neutral in dem Sinne ist, daß er sowohl im Interesse von Schülern, auch Eltern und Lehrern, eingesetzt werden kann, als auch gegen sie. Hier kann nur das Berufsethos des Anwenders entscheiden.

Das Verfahren selbst kann als Instrument zu optimalen Diskrimination heterogener Merkmalsträger definiert werden. Dabei wird nach SKARABIS (1970, S. 1)

"unter DISKRIMINANZ ... die Zuweisung eines Individuums in eine von mehreren bekannten Teilgesamtheiten verstanden, wenn das Individuum Träger einer p-dimensionalen Beobachtung ist, d.h. an ihm p Merkmale beobachtet sind."

Unsere Merkmalsträger sind die Schüler, die Merkmale sind Testleistungen und die Teilgesamtheiten sind die verschiedenen Schularten.

Oder um KRAUTH & LIENERT (1973, S. 16f) zu zitieren, die zwischen

- POSTHOC-Klassifikation, zu denen sie die Faktorenanalyse (Typenanalyse) zählen, und
- ANTHAC-Klassifikationen, zu denen sie die Diskriminanzanalyse zählen, unterscheiden.

Sie definieren:

"Die Anthac-Klassifikation verfolgt das Ziel, Merkmalsträger (Patienten) Klassen zuzuordnen, die vor der Erhebung des multivariaten Datenmaterials bereits existieren oder durch ein von den beobachteten Merkmalen unabhängiges Kriteriumsmerkmal definiert werden."

Unseres Erachtens ist diese Einteilung seit der Entwicklung der Automatischen Klassifikation (AUKL) von Faber & NOLLAU (1969) nicht mehr gerechtfertigt, da sie ebenfalls geeignet ist, Posthoc-Klassifikationen zu erlauben.

Es soll hier keine Ableitung der Diskriminanzanalyse erfolgen. Dazu verweisen wir auf HOLM (1979), HOPE (1975), BAUMANN (1971) und ÜBERLA (²1971) und die in diesem Kontext bereits erwähnten Autoren. Wir fügen lediglich im Anhang die Formeln auf, die unserem Algorithmus zugrunde liegen.

9.2 Diskriminanzanalyse der HS,
 RS, OS über 7 PMA 9-10

Zunächst sei nochmals festgehalten, daß weder diese noch andere Tests (u.W.) irgendwie in die Frage involviert waren, welche weiterführende Schulen unsere Probanden besuchen sollten. Daher sei hier zunächst eine D-Analyse über die Tests vorgenommen, die zum Zeitpunkt des Übergangs von der Grundschule zu einer weiterführenden Schule erhoben wurden, die PMA 9-10, als also alle Kinder abgesehen von Zurückstellung und Sitzenbleiben, die gleiche Zahl von Schuljahren durchlaufen und abgesehen von den SS, die die damalige Volksschule besucht hatten.

Wir "überprüfen" gewissermaßen die Richtigkeit der damals - von wem auch immer; Eltern, Lehrer, Schüler - getroffenen Entscheidung. Dabei gehen wir von der Annahme aus, daß die getroffenen Entscheidungen im großen und ganzen richtig waren, d.h. nur wenige Schüler, auf wessen Veranlassung auch immer, ihre Entscheidung später geändert haben. Diese Annahme scheint uns notwendig, da andernfalls das Entscheidungssystem nicht akzeptabel wäre. Ziel jedes Entscheidungssystems muß es ein, optimale Entscheidungen zu fällen, d.h. Fehlurteile zu minimieren. Hinzu kommt hier, daß die Entscheidungen helfen sollten, die individuelle freie Entfaltung

der Persönlichkeit zu gestatten und zu fördern.
Daß dies vom Individuum am besten auf möglichst
hohem Niveau reflektiert werden kann, versteht
sich u.E. von selbst.

Die Gruppe der SS (= 12) wurde hier außer acht ge-
lassen, da wir in einigen Vorläufen gesehen hatten,
daß die Probanden eine homogene Gruppe bildeten,
abgesehen von Pb 68, einem älteren, 1960 eingeschul-
ten Mädchen, das später zur Sonderschule überwiesen
worden war. Der Vater gehörte zur Gruppe "ausführende
Tätigkeiten ohne Berufsausbildung". Bei Diskrimi-
nanzanalysen über Tests der Untersuchung 1964 wurde
sie als einzige zuweilen (je nach Tests) als Haupt-
schülerin klassifiziert.

Bei der Darstellung der Diskriminanzanalysen (D-
Analysen) gehen wir immer so vor, daß wir in den
Tabellen zeilenweise die Ausgangsbedingungen und
spaltenweise die Zuordnungen nach den Diskriminanz-
analysen wiedergeben.

Für die Berechnung der D-Analyse über die 7 PMA 9-10
waren die drei Gruppen HS, RS und QS vorgegeben.
Es sollte nun mit Hilfe der D-Analyse geprüft werden,
wie ähnlich die kognitiven Strukturen, repräsentiert
durch die PMA 9-10, der Schüler einander seien.

Formal wird dabei so vorgegangen, daß für jede der drei Gruppen getrennte Diskriminanz-Funktionen berechnet werden. Danach werden für jeden Probanden alle (hier: drei) Funktionen ausgerechnet, und angegeben, welcher Gruppe er mit größter Wahrscheinlichkeit am ähnlichsten ist. So kann also z.B. die kognitive Struktur eines HS derjenigen der OS ähnlicher sein als derjenigen der übrigen HS. Oder ein OS könnte den RS ähnlicher sein.

Angewandt auf Beratungsfragen (z.B. Übergang zur ...) könnte dies bedeuten, daß in unseren beiden Beispielen der HS besser zur OS und der OS besser zur RS gegangen wäre, vorausgesetzt, und dies ist essentiell, die kognitiven Leistungen wären das alleine relevante Kriterium für solch eine Entscheidung. Wir hätten in die Analysen gern Motivations- und Interessenvariablen einbezogen. Leider gaben unsere Instrumente dazu jedoch nichts her.

Wir haben die generellen Anmerkungen vorweg erwähnt, weil sie für alle folgenden Analysen präter propter gelten. Es sollte auch immer präsent sein, daß für echte Entscheidungen das Verfahren leicht (nicht prinzipiell) modifiziert werden müßte (vgl. FABER & NOLLAU 1969a, 1969b; BAUMANN 1971). Der mögliche Einwand, daß hier die Gruppen bereits vorgegeben sind, trifft die Leistungsfähigkeit des Verfahrens nicht.

Doch nun zu Tabelle 38 mit den Diskriminanzanalyse-Ergebnissen.

Tabelle 38: Zuordnung nach Diskriminanzanalyse, 7 PMA 9-10, Stichproben HS, RS, OS.

$$D^2 = 179,1 \quad df = 14$$

Zuordnung nach D-Analyse +)				
	HD	RD	OD	Su.
HS	73/73%	20/20%	7/7%	100
	38-99%	39-63%	47-85%	
RS	15/30%	17/34%	18/36%	50
	39-78%	39-61%	41-71%	
OS	6/10%	18/29%	38/61%	62
	35-67%	40-62%	36-91%	
Su.	94/44%	55/26%	63/30%	212

Wir wollen diese erste Tabelle etwas ausführlicher besprechen, um uns bei den folgenden jeweils auf das Wesentliche beschränken zu können.

Das MAHALANOBIS-D-Quadrat (im folgenden D^2 genannt) ist mit 179,1 hoch signifikant, zeigt also, daß die Gruppen verschiedene kognitive Strukturmuster haben.

+) Bei den Tabellen der Diskriminanzanalysen sind spaltenweise immer die Zuordnungen gemäß den D-Analysen eingetragen. Es bedeuten: HD=Hauptschüler nach D-Analyse, RD=Realschüler nach D-Analyse und OD=Oberschüler nach D-Analyse.

In jeder Zeile ist außerdem noch eingetragen, mit welcher Wahrscheinlichkeit die D-Analyse die Probanden dieser Zeile zugeordnet hat.

Beispiel: Die D-Analyse hat 7 HS identifiziert, die den Oberschülern ähnlicher sind als den Hauptschülern. Diese größere Ähnlichkeit zu den Oberschülern schwankt für die sieben Schüler zwischen 47-85%.

Wie können wir nun die Güte dieser Diskrimination beurteilen? Nun, wir kennen das Schulschicksal aller OS, dank der mühevollen Arbeit von FIEGENBAUM. Außerdem kennen wir von den RS den Schulstatus von 1969 und den der HS von 1968, dem Zeitpunkt als die ersten HS ihre allgemeine Schulpflicht bereits erfüllt hatten.

Die Tabelle 39 schlüsselt die Zeile OS der Tabelle 38 auf.

Tabelle 39: Auffächerung der OS aus Tabelle 38

	Zuordnung nach D-Analyse			
	HD	RD	OD	Su.
OS, Abi 72	0	1	20	21
OS, Abi 73	0	4	5	9
OS, Abi 74	2	2	2	7
OS, MR (mittlere Reife)	1	7	8	16
OS, DD (Drop Out)	3	3	3	9
Su.	6	18	38	62

Danach können wir folgendes feststellen:

$20 + 5 + 2 = 27$ Abiturienten wurden richtig klassifiziert. Ferner $1 + 7 = 8$ MR und $3 + 3 = 6$ Drop Outs. Für die MR und DO gilt das Wort "richtig" insofern, als ihre kognitive Struktur nicht den OS entsprach. Hier hätte vielleicht - besonders bei den DO - eine andere Bildungsempfehlung getroffen werden müssen. Dies gilt präter propter auch für die verspäteten Abiturienten.

Ganz sicher wären aber mit Hilfe derartiger Analysen die Notwendigkeit einer Beratung bei den acht OS, die "nur" die mittlere Reife geschafft haben, und die drei OS, die als DO die Schule ohne jeden Abschluß verlassen mußten, deutlich geworden. Die einzige richtige Fehlzuordnung ist der den RD zugerechnete OS, der jedoch ohne Verzögerung sein Abitur geschafft hat. Es handelt sich hierbei um einen älteren Jungen, dessen Vater eine Führungsposition auf Meisterebene innehat.

Die Tabelle 4G schlüsselt die Zeile PS der Tabelle 38 auf.

Tabelle 40: Auffächerung der RS aus Tabelle 38

	Zuordnung nach D-Analyse			
	HD	RD	OD	Su.
RS, Klasse 11	11	12	14	37
RS, Klasse 10	2	3	1	6
RS, Klasse 9	1	1	1	3
RS, Klasse ?	1	1	2	4
Su.	15	17	18	50

Wie schon bei der Stichprobenbeschreibung sichtbar wurde, ist die Gruppe der RS die am wenigsten homogene Gruppe. Daß jedoch die größte Gruppe der kognitiv-strukturellen OD bereits die mittlere Reife (MR) erlangt hat, deutet an, daß zumindest diese 14 Schüler auch ein Gymnasium hätten besuchen können.

Die Tabelle 41 schlüsselt die Zeile HS der Tabelle 38 auf.

Tabelle 41: Auffächerung der HS aus Tabelle 38

	Zuordnung nach D-Analyse			
	HD	RD	OD	Su.
HS, Klasse 9	36	17	5	58
HS, Klasse 8	32	3	2	37
HS, Klasse 7	5	0	0	5
Su.	73	20	7	100

Analog zu den RS läßt sich für die HS sagen, daß zumindest die kognitiv-strukturellen siebzehn RD und fünf OD der Klasse 9 (Hauptschulabschluß) eine Real- bzw. Oberschule hätten besuchen können. Für die Brauchbarkeit dieses Verfahrens ist auch die Tatsache ein Hinweis, daß alle HS der Klasse 7 den HD zugeordnet werden, sie also eine kognitive HD-Struktur aufweisen.

In dieser Analyse hat sich gezeigt, daß die Schülerfrequenzen der einzelnen Schulen sich im Ernstfall gar nicht beträchtlich geändert hätten.

Um herauszufinden, welchen Sozialschichten (Beruf des Vaters) die Unter- und Überforderten angehören, haben wir sie entsprechend zusammengefaßt. Die so entstandene 9-Feldertafel, Tabelle 42, ist auf dem 5%-Niveau signifikant, d.h. aus den Schichten I+II und III werden erheblich mehr Schüler auf Schulen geschickt, die sie überfordern, bzw. aus den Schichten IV+V entstammen erheblich weniger Überforderte, als statistisch zu erwarten wäre. Bei den Unterforderten sind es die Schüler der Gruppe I+II, deren Unterforderungsrate erwartungswidrig niedrig ist. Die Signifikanz wird also hauptsächlich - wenn die kognitive Struktur ausschlaggebend gewesen wäre für die besuchte Schule - durch die "falsche" Schulwahl der Überforderten

der oberen Schichten verursacht, wer immer die Entscheidung getroffen habe (vgl. KEMMLER 1976).

Tabelle 42 : Kognitive Struktur nach 7 PMA 9-10, Sozialschicht und Schulbesuch

	Beruf des Vaters (Schicht)						Su.
	I+II		III		IV+V		
	o	e	o	e	o	e	
Unterforderte	6	10	14	12	23	20	43 ⁺
Überforderte	13	9	15	11	10	18	38
Adäquate	30	33	35	41	64	68	129
	49		64		97		210

o := beobachtet, e := erwartet, $p < 5\%$

⁺ hier fehlen 2 Berufsangaben

Vergessen darf man bei unserer Stichprobe aus und um Münster nicht, daß Münster neben Bonn schon in den 60er Jahren zu den Städten mit dem am besten ausgebauten Bildungssystem gehörte, und von daher die Quoten der RS und OS schon damals erheblich über dem Bundesdurchschnitt lagen.

9.3 Diskriminanzanalyse der HS, RS, OS über 6 PMA 13-14

Mit dieser Diskriminanzanalyse, vgl. Tabelle 43, soll zunächst festgestellt werden, ob sich bedeutsame Veränderungen hinsichtlich der Häufigkeiten der Zu-

ordnung ergeben haben (das $D^2 = 251,4$ ist hoch signifikant), ob also z.B. mehr Oberschüler als bei der PMA 9-10 D-Analyse etwa zur Realschule oder Hauptschule hätten gehen sollen, oder ob die Zahl der Schüler mit Oberschulstruktur gegenüber 1964 sogar noch angestiegen ist.

Tabelle 43 : Zuordnung nach Diskriminanzanalyse,
6 PMA 13-14, Stichprobe HS, RS, OS.

$$D^2 = 251,4 \quad df = 12$$

Zuordnung nach D-Analyse				
	HD	RD	OD	Su.
HS	81/81%	14/14%	5/5%	100
	40-99%	40-76%	45-79%	
RS	8/16%	26/52%	16/32%	50
	46-83%	44-78%	43-83%	
OS	5/8%	15/24%	42/68%	62
	40-80%	41-75%	36-95%	
Su.	94/44%	55/26%	63/30%	212

Auf der Diagonalen finden sich die Zuordnungen, die den jeweiligen Schul-Pattern entsprechen, also $81 \text{ HS} + 26 \text{ RS} + 42 \text{ OS} = 149$ Schüler, entsprechend 70% d.h. eine größere Zahl (5% mehr als 1964) haben sich der Struktur ihrer Schule angepaßt. Oder hat sich die Schule ihnen angepaßt? Die Zahl der Überforderten

8 RS + 15 OS + 5 OS = 28 Schüler ist um 11 Schüler gesunken. Die Zahl der Unterforderten ist ebenfalls um 10 Schüler gesunken. Man kann also vermuten, daß alle drei Schulen eine Anpassung zum Durchschnitt vollziehen.

Diese Frequenzvergleiche geben aber noch keine Auskunft darüber, wie sich möglicherweise die Intelligenzstrukturen der einzelnen Schüler von 10-14 Jahren verändert haben, einem Alter, in dem noch stärkeres intellektuelles Wachstum angenommen wird. Üblicherweise gewinnt man für die ganze Gruppe derartige Hinweise über Korrelationskoeffizienten zwischen Tests. Wir haben hier einen Weg gewählt, der den einzelnen Schüler etwas mehr ins Blickfeld bringt. Dazu haben wir die Ergebnisse der beiden gerade besprochenen D-Analysen der Tabellen 38 und 43 zusammengetragen und in Tabelle 44 vereinigt.

Tabelle 44: Zusammenfassung der Zuordnungen der beiden Diskriminanzanalysen, 7 PMA 9-10, 6 PMA 13-14

Zuordnung nach <u>beiden</u> D-Analysen ⁺⁾							
	HD/HD	RD/RD	OD/OD	HD/RD	RD/OD	HD/OD	Su.
HS	62	4	1	25	3	5	100
RS	4	12	12	12	7	3	50
OS	3	10	34	3	10	2	62
Su.	69	26	47	40	20	10	212

⁺⁾ HD/HD bedeutet: Alle Schüler dieser Spalte wurden in beiden D-Analysen als Hauptschüler klassifiziert. Die übrigen Kombinationen analog.

Die Zusammenfassung erfolgte folgendermaßen:
Wir haben jeweils die Schüler zu Gruppen, entsprechend ihrer jeweiligen Zuordnung, zusammengefaßt. Also die vier RS unter HD/HD werden in beiden D-Analysen als Hauptschüler klassifiziert. Ebenso wurden die drei HS unter RD/OD in allen drei Fällen einmal zu Realschülern und einmal zu Oberschülern klassifiziert. Offen bleibt hierbei, in welchem Verfahren welche Klassifikation erfolgte.

Nun zur Tabelle selbst: $69 + 26 + 47 = 142$ (= 67%), letzte Zeile, wurden in beiden D-Analysen in gleicher Weise zugeordnet oder besser: klassifiziert. In weiteren $40 + 20 = 60$ (= 28%) erfolgt die Klassifikation benachbarter Gruppen. Und nur bei 10 Schülern (= 5%) besteht hier die maximal mögliche Diskrepanz von zwei Gruppen. Eine derartig stabile Klassifikation hätten wir eigentlich über den Zeitraum von 4 Jahren und bei dem Alter der Schüler nicht erwartet. Wir deuten diesen Befund unter Einbeziehung aller Unzulänglichkeiten der Untersuchung als ein starkes Indiz dafür, daß intellektuelle Förderung in jedem Falle weit vor dem 10. Lebensjahr einsetzen sollte. Und im Unterschied zu den amerikanischen Head-Start-Projekten muß sie vermutlich auch über die gesamte Schulzeit andauern. Ein simpler Vergleich dazu sei gestattet: Wer Sprachen lernt oder Leistungssport betreibt, muß auch ständig im Training bleiben, sollen seine Fähig-

keiten nicht deutlich absinken. Es ist u.E. noch nicht nachgewiesen, daß dies im kognitiven Bereich nicht allgemein gilt.

Die am meisten voneinander abweichenden Klassifikationen bilden die Schüler HD/OD, letzte Spalte der Tabelle 44; d.h. sie wurden in einer D-Analyse als Hauptschüler, in der anderen als Oberschüler klassifiziert. Sie sollen deshalb etwas genauer beschrieben werden.

Zunächst die fünf echten Hauptschüler:

- Pb 13: jüngerer Junge, einmal sitzengeblieben, GSW 64 = 51, GSW 68 = 56, Berufsgruppe des Vaters (=BdV) "Stehkragenberufe"
- Pb 264: älterer Junge, 9. Klasse, GSW 64 = 50, GSW 68 = 47, BdV unbekannt
- Pb 57: älterer Junge, 9. Klasse, GSW 64 = 52, GSW 68 = 57, BdV "manuelle Berufe"
- Pb 76: jüngerer Mädchen, 9. Klasse, GSW 64 = 62, GSW 68 = 58, BdV "Stehkragenberufe"
- Pb 151: jüngerer Mädchen, 9. Klasse, GSW 64 = 53, GSW 68 = 60, BdV "manuelle Berufe"

Als nächstes die drei echten Realschüler:

- Pb 144: jüngerer Junge, 8. Klasse, Übergang zur RS nach Klasse 5, GSW 64 = 58, GSW 68 = 48, BdV "Stehkragenberufe"
- Pb 285: jüngerer Mädchen, Zurückgestellt, 7. Klasse, Übergang zur RS nach Klasse 5, GSW 64 = 53, GSW = 58, BdV "Führungspos. auf mittlerer Ebene"
- Pb 15: älteres Mädchen, 8. Klasse, Übergang zur RS nach Klasse 5, GSW 64 = 46, GSW 68 = 52, BdV "manuelle Berufe"

Und schließlich die beiden echten Oberschüler:

- Pb 320: jüngerer Junge, Zurückgestellter, Abitur 73, GSW 64 = 57, GSW 68 = 53, BdV "Stehkragenberufe"
- Pb 272: jüngerer Mädchen, Zurückgestellte, Abschluß Mittlere Reife, GSW 64 = 50, GSW 68 = 53, BdV "Stehkragenberufe"

Es fällt auf, daß 50% der Väter dieser Schüler der Gruppe der "Stehkragenberufe" zugerechnet werden. Von daher ergibt sich vielleicht ein Hinweis für die Klassifikationsschwankungen, da ja in der "Mittelschicht" die Bildungsaspirationen besonders ausgeprägt sind.

9.4 Diskriminanzanalyse der HS,
RS, OS über GSW 1964

Wie bekannt, wird, wenn bei Schullaufbahnberatungen auch Intelligenztestergebnisse hinzugezogen werden, in aller Regel letztlich der IQ Teil der Entscheidungsfindung. Wir haben den Vorgang für unsere Stichprobe nachzuvollziehen versucht. Angenommen wird für die einzelnen Schulen ein Eintritts-Schwellenwert (cutting point bei LIENERT² 1967, S. 479; cut-off points bei HÄRNQUIST 1968, S. 74), d.h. einen Mindest-IQ für die jeweilige Schulart.

Wir haben nun über den GSW 1964 eine D-Analyse berechnet, deren Ergebnisse Tabelle 45 zeigt. Die beobachteten Frequenzen weichen nicht alszu sehr von denen der Tabelle 43 ab.

Tabelle 45 : Zuordnung nach Diskriminanzanalyse
GSW 1964, Stichproben HS, RS, OS

$$D^2 = 124,5 \quad df = 2$$

Zuordnung nach D-Analyse				
	HD	RD	OD	Su.
HS	76	18	6	100
RS	19	14	17	50
OS	6	21	35	62
Su.	101	53	58	212

Zudem haben wir aber nach den errechneten Zuordnungswahrscheinlichkeiten für jede Gruppe die Grenzwerte in Tabelle 46 zusammengefaßt. Es ist deutlich zu erkennen, daß sich die GSW 1964 für HS, RS und OS beträchtlich überlappen.

Tabelle 46 : Grenzwerte des GSW 1964
nach Diskriminanzanalyse

SOLL-Lage nach D-Analyse				
	HD	RD	OD	Streubreite
IST- LAGE	HS 31-50	54-54	58-77	31-77
	RS 37-50	54-54	58-64	37-64
	OS 45-48	52-55	57-76	45-76
	31-50	52-55	57-77	31-77

Intelligenz, wie sie mit diesen oder ähnlichen Tests erfaßt wird, kann für unsere Stichprobe bei der

Entscheidung, welche Schule jemand besucht, kaum eine Rolle gespielt haben. Ob auf diese Weise die Ziele, wie sie der Deutsche Bildungsrat 1970 in seinem "Strukturplan" formuliert hat, realisiert werden können, kann bezweifelt werden, nämlich:

"Die Ansprüche auf schulische Bildung und freie Entfaltung der Persönlichkeit führen zu dem Grundsatz, daß jeder einzelne so weit wie möglich zu fördern ist" (Strukturplan, 1970, S. 31).

und

"Gleichheit der Chancen wird in manchen Fällen nur durch die Gewährung besonderer Chancen zu erreichen sein" (Strukturplan, 1970, S. 30).

9.5 Weitere Diskriminanzanalysen

Wir haben noch eine ganze Anzahl weiterer Diskriminanzanalysen durchgerechnet, z.B.

- Junge/Mädchen,
- Ältere/Jüngere,
- Altersgemäß Eingeschulte/Zurückgestellte.

In jedem Falle erhielten wir sehr signifikante D^2 -Werte, was auf unterschiedliche kognitive Strukturen in den miteinander verglichenen Gruppen hinweist. Auf die tabellarische Darstellung der Ergebnisse kann u.E. verzichtet werden. Im folgenden wollen wir jedoch zwei interessante D-Analysen besprechen, die jede für sich, zwei Effekte, einen Geschlechts- und einen schulartspezifischen, in sich vereinigen.

Darüber hinaus kann man feststellen, daß jeweils die größte Gruppe der eigenen Schulart entsprechend klassifiziert wird, das sind 74 HS, 26 RS und 37 OS, zusammen 137 Schüler oder 65%. Noch interessanter erscheinen die Strukturen der bisher unberücksichtigt gebliebenen 75 Schüler. Von den verbleibenden 43 Jungen dieser Gruppe werden 28 der Jungengruppe einer anderen Schulart zugeordnet und nur 15 Jungen ähneln in ihrer Intelligenz-Struktur eher Mädchen einer anderen Schulart, also eine Relation von annähernd 2 : 1. Ähnlich ist es bei den Mädchen. von 32 Mädchen werden 23 wieder nach den Mädchen einer anderen Schulart klassifiziert und nur 9 nach den Jungen anderer Schularten, hier ein Verhältnis von 2,6 : 1. Das Ergebnis ist ein sicheres Indiz für die in unserer Stichprobe unterschiedlichen Intelligenz-Strukturen, sowohl hinsichtlich der Schulart als auch hinsichtlich des Geschlechts, sagt aber nichts über das Zustandekommen aus.

9.7

Diskriminanzanalyse der
Schüler nach Schulart und Geschlecht,
über 11 Leistungstests 1968

Die gleiche D-Analyse haben wir auch mit den 11 Leistungstests der Untersuchung 1968 durchgeführt, weil wir wissen wollten, ob infolge des Unterrichts in den verschiedenen weiterführenden Schularten

sich die Struktur-Muster "gelockert" haben würden.
Tabelle 48 zeigt die Klassifikationen der D-Analyse, deren $D^2 = 445,9$ hochsignifikant ist.

Tabelle 48 : Zuordnung nach Diskriminanzanalyse
11 Leistungstests von 1968
Stichprobe nach Schultart und Geschlecht
 $D^2 = 445,9$ $df = 50$

Zuordnung nach D-Analyse							
	HDj	HDm	RDj	RDm	ODj	ODm	Su.
HSj	30	13	3	1	4	1	52
HSm	12	29	3	3	1	0	48
RSj	0	0	15	1	4	1	21
RSm	2	3	1	16	2	5	29
OSj	3	1	9	6	23	0	42
OSm	2	0	0	4	2	12	20
Su.	49	46	31	31	36	19	212

j = Junge m = Mädchen

Auf der Diagonalen finden sich 125 Schüler (= 59%), deren Zuordnung der Realität entspricht. Außerhalb der Klassifikation, die der eigenen Schultart unabhängig vom eigenen Geschlecht entspricht, gibt es unter 33 Jungen dreiundzwanzig, deren Leistungsstruktur derjenigen von Jungen anderer Schultarten entspricht. Umgekehrt haben 10 Jungen Leistungsstrukturen, die denen von Mädchen gleichen, insge-

samt ein Verhältnis von 2,3 : 1. Von den 25 Mädchen wurden 15 wieder zu den Mädchen klassifiziert und ebenfalls 10 zu den Jungen anderer Schularten. Hier ist die Relation 1,5 : 1.

Zusammenfassend können wir also feststellen, daß die Diskriminanzanalyse eine Methode ist, die sich prinzipiell zu Diagnose- und auch Prognosezwecken eignet. Besonders mit Hilfe von programmierbaren Tischrechnern mit einer Speicherkapazität von ca. 32K scheint uns diese Prozedur zu den Methoden der Wahl für derartige Problemstellungen zu gehören, wie sie z.B. AURIN (1968); HELLER (1970); BETHÄUSER & REICHENBECHER (1976); HELLER (1976); und HELLER/ROSEMAN/STEFFENS (1978) im Kontext der Schullaufbahnberatung (Klassifikation und Prädiktion) angewendet bzw. problematisiert haben. Voraussetzung für den Einsatz sind und bleiben aber in erster Linie valide Tests.

Abschließend sei noch angemerkt, daß, wenn wir nur die kognitiven Strukturen, wie sie von uns erfaßt worden sind, zugrunde legen, noch ein erhebliches Potential dessen, was gemeinhin mit "Begabungsreserve" umschrieben wird, vorhanden ist. Sie brauchte bloß gefördert zu werden.

10.0 ZUM UMWELTEINFLUSS BEI DER INTELLIGENZENTWICKLUNG

Vorbemerkung: Mit den Arbeiten von JENSEN (1969, 1972, 1973), EYSENCK (1971, 1972, 1973), HERNSTEIN (1971) und JENCKS (1972), um nur einige zu nennen, ist eine Diskussion neu - und offensichtlich heftiger als je zuvor - entbrannt, die in dieser Form der Problemstellung seit langem als unergiebig und müßig bezeichnet worden war (vgl. ANASTASI 1958, dt. 1973). Gemeint ist die Diskussion auch als IQ-Kontroverse bekannt - über die Beeinflussung des Verhaltens durch Erbe und Umwelt, hier speziell der Anteile, die Erbanlagen einerseits und Umweltwirkungen andererseits bei der Intelligenz-entwicklung zuzuschreiben seien. Allein JENSEN führt in seinem 1972 erschienenen Buch "Genetics & Education" mehr als 100 Titel auf, die sich nur auf seinen 1969 erschienenen Aufsatz über das Ausmaß der Steigerungsfähigkeit von Intelligenz und Schulleistung beziehen. In der Bundesrepublik Deutschland sind Darstellungen und Bewertungen in der Z. f. Pädagogik (1970, Heft 2), in einem Sammelband von SKOWRONEK (1973), ROTH/DAUMEN-LANG (1972), ZIMMER (1975), KNOCHKE (1977), SCHMID (1977) und MERZ/STELZL (1977) erfolgt, auf die hier verwiesen werden kann.

Abgesehen von der trivialen, Überall vorfindbaren Auffassung, wonsch Erbe und Umwelt die Intelligenzent-

wicklung beeinflussen, reichen die Urteile (international) von schroffer Ablehnung bis (modifizierter) Zustimmung, wenn eine Quantifizierung der jeweiligen Anteile als Varianzkomponenten vorgenommen wird, die nach JENSEN und anderen Autoren für die Erbanlagen einen Betrag von 45 - 90 Prozent ergibt. Mithin verblieben für Umweltwirkungen 10 - 55 Prozent der Gesamtvarianz.

In dieser Kontroverse kann jedoch nicht übersehen werden, daß wegen der eminent politischen Bedeutung der pädagogischen Konsequenzen - wie immer die Ergebnisse auch ausfallen mögen - je nach Standpunkt des Autors Forschungsergebnisse mehr oder weniger stark selektiv rezipiert werden. So weist z.B. JENSEN mehrfach auf Unzulänglichkeiten seiner Daten hin, was ihn aber nicht hindert, seine Schätzung der Erbvarianz derart massiv vorzutragen und darauf zu insistieren, daß etwaige Relativierungen offensichtlich von Kritikern nicht mehr ernst genommen oder gar als Verschleiierungstaktik charakterisiert werden. In der Tat scheint es jedoch angebracht, angefangen von den unterschiedlichen verwendeten Tests, den Stichproben, die in ihrer Alters- und Sozialstruktur mehrere Generationen umfassen, der verschiedenen ethnischen Herkunft der Daten, bis hin zu den unreichend beschriebenen Umwelten, in denen die

Probanden aufgewachsen sind, skeptisch zu sein, was die Brauchbarkeit bisheriger Ergebnisse anlangt (s. JENSEN 1969, S. 43). Andererseits muß die Aktualisierung dieser Problematik angesichts steigender Jugendarbeitslosigkeit, Verknappung von Ausbildungsplätzen, Verschärfung der Hochschulzugangsbedingungen, etc. die Forschung stimulieren, sofern auf politischer Ebene endlich eingesehen wird, daß das Unbehagen in der Gesellschaft, das es jetzt zu Überwinden gilt, nicht zuletzt eine Folge jahrzehntelanger (relativer) Vernachlässigung pädagogisch-psychologischer Forschungsfinanzierung und -förderung ist. "Notprogramme" helfen nur über den Tag, lösen aber letztlich keines der anstehenden Probleme. Wenn man ferner berücksichtigt, daß die Schwierigkeiten zunächst noch massiv ansteigen und erst Mitte der 80er Jahre allmählich geringer werden - und dies primär aufgrund sinkender Geburtsjahrgangsstärken - so bleibt keine Zeit zu verlieren, Grundlagenforschung in diesem Bereich in großem Umfang zu fördern, die unabhängig von parteipolitischen Egoismen durchgeführt werden kann.

Man könnte folgende These formulieren:

Wissenschaft, Politik und Gesellschaft müssen der Erbe-/Umweltproblematik der Intelligenz um so mehr Beachtung

beimessen, je höher der Erbanteil zu veranschlagen ist, weil nur dann rechtzeitig gezielte Anstrengungen unternommen werden können, einem sonst zumindest nicht auszuschließenden Sozialdarwinismus zu entgehen, wie er bereits propagiert wird (vgl. HERRNSTEIN 1973, dt. 1974; HUSEN 1975, S. 24). Es erscheint als eine unangemessene Erwartung, Intelligenz sei für kommende Generationen von geringerer Bedeutung als heute, etwa zugunsten von mehr Solidarität in einer humaneren Gesellschaft. Während sogenannte Industrieländer noch Berichte wie "Born to fail" (WEDGE/PROSSER 1973) verfassen - wobei hier nicht gegen sorgfältige Bestandsaufnahmen argumentiert werden soll - hat ein Land wie Venezuela ganz andere Konsequenzen gezogen: Dort wurde 1979 ein "Ministry of State for the Development of Intelligence" gegründet (STERNBERG 1981; MACHADO 1981a, 1981b), das erste und einzige Ministerium auf der Welt - soweit bekannt - dessen einzige Aufgabe es ist, die kognitiven Fähigkeiten aller Bürger zur Entfaltung zu bringen. Hier scheint THOMAS v. AQUIN eine späte Bestätigung zu erfahren, für den die Entwicklung einer humaneren Gesellschaft geradezu der Intelligenz bedarf.

Da - wie schon früher erwähnt - die KEMMLERSche Längsschnittuntersuchung kein Experiment im üblichen Wort-sinn ist (vgl. METZGER 1952), sondern der Erkundung von Zusammenhängen dient, die zu präzisen Hypothesenprüfungen in anzuschließenden Experimenten führen sollen, lag die Frage nahe, ob die Längsschnittuntersuchung Informationen zur Erbe-/Umweltproblematik würde liefern können. Dies schien zunächst nicht möglich, da für diesen Forschungsbereich die Zwilling- und Familienforschung von besonderer Bedeutung sind und die Längsschnittuntersuchung - bezogen auf Erbanlagen für die Intelligenz - aus genotypischen nicht verwandten, also unabhängigen Individuen besteht.

Eine empirische Überprüfung bisheriger Aussagen zu diesem Thema ist also mit Daten der Längsschnittuntersuchung nur dann möglich, sofern dazu Daten von nicht-verwandten Probanden zur Varianzschätzung herangezogen wurden, was bei JENSEN der Fall ist.

10.1 Schätzung der Umweltvarianz der Intelligenz nach JENSEN

Für die Schätzung der phänotypischen Varianz (V_P) der Intelligenz innerhalb einer Population verwendet JENSEN (und andere) die Formel:

$$V_P = V_G + V_{AM} + V_D + V_i + V_E + 2COV_{HE} + V_I + V_e = 100\%$$

Es bedeuten:

V_P := phänotypische Varianz der Population

V_G := genetische (oder additive) Varianz

V_{AM} := Varianz, die auf gezielte Partnerwahl zurückgeht. $V_{AM} = 0$ unter Panmixie-Bedingungen.

V_D := Dominanzabweichungsvarianz

V_I := Epistase (Interaktion zwischen Genen an zwei oder mehr loci)

V_E := Umweltvarianz

COV_{HE} := Kovarianz von Erbe und Umwelt

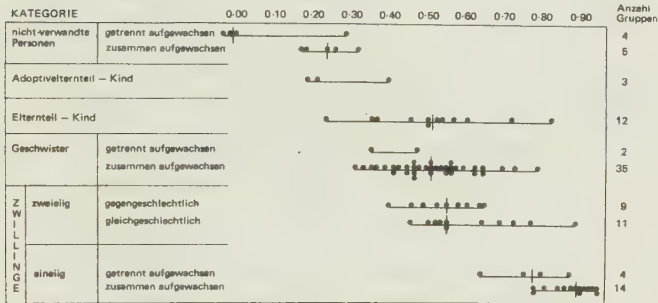
V_I := echte statistische Interaktion zwischen Erb- und Umweltfaktoren

V_e := Meßfehler (mangelnde Zuverlässigkeit)

Der Gesamteinfluß der Umwelt ergibt sich also aus den Komponenten V_E , COV_{HE} und V_I und darauf sollen sich auch unsere Intentionen in erster Linie beziehen. Dieses Modell impliziert die Additivität der Varianzkomponenten. es gibt jedoch Untersuchungen, die gezeigt haben, daß stärkere Interaktionen als bisher angenommen zwischen Erbe und Umwelt andere Modellvorstellungen erfordern (nach ANASTASI, deutsch 1973, S. 9).

Bevor wir auf die Schätzung der Umweltvarianz näher eingehen, soll eine Tabelle von ERLÉNMEYER-KIMLING und JARVIK (1963) wiedergegeben werden, die JENSEN - ergänzt um neuere Werte, wobei er jedoch nur auf BURT (1966) verweist - zur Ermittlung seiner Schätzwerte zugrunde legt; nicht zuletzt, weil die Streubreite der Korrelationen in den verschiedenen Gruppen einen genaueren Überblick über die empirischen Fakten gibt, als dies nur die Medianwerte vermögen (JENSEN's Tab. 2).

Tabelle 49: Korrelationskoeffizienten von Intelligenztestresultaten nach ERLÉNMEYER-KIMLING und JARVIK (1963) bzw. JENSEN (1969) (hier zitiert nach SCHMID (1977))



Es fällt auf, daß etwa bei Untersuchungen zwischen Eltern/Kind, zusammenaufgewachsenen Geschwistern und gleichgeschlechtlichen Zwillingen Korrelationen gefunden wurden, die in ihren oberen Bereichen den Median erreichen, ja teilweise gar das ganze Spektrum der Korrelationen für getrennt aufgewachsene monozygote Zwillinge (EZ) überlappen, bis hin zum Median für zusammen aufgewachsene EZ. Bei der gesellschaftspolitisch-pädagogischen Bedeutung, die JENSEN seinen Ergebnissen zumißt und/oder von anderer Seite zugemessen wird, und der Vielfalt der insgesamt angewendeten statistischen Analysetechniken hätte eigentlich ein detaillierter Erklärungsversuch (anhand der Originaldaten) für diese teilweise signifikanten Abweichungen vom zugehörigen Median erwartet werden können. JENSEN jedoch begnügt sich mit der Wiedergabe der Feststellung von ERLÉNMEYER-

KIMLING und JARVIK, in die Zusammenstellung seien insgesamt 30 000 (!) Korrelationspaare aus 8 Ländern in 4 Kontinenten über mehr als zwei Generationen eingegangen, und des Erstaunens der beiden Autoren, wie gut doch bei aller Heterogenität der Herkunft und des Zustandekommens der Daten die empirischen Mediane mit der Modellannahme (Polygenie), die üblicherweise zur Erklärung erblicher Unterschiede geistiger Fähigkeiten verwendet wird, kompatibel seien und weist auf die relativ konstante Differenz der Korrelationskoeffizienten zwischen getrennt und zusammen aufgewachsenen Probanden bestimmter Verwandtschaftsgrade hin, was die Gleichartigkeit der Umweltwirkungen demonstrieren soll. Uns jedenfalls scheinen JENSEN's Anmerkungen in diesem Punkt völlig unzureichend.

Tabelle 50 ist nun JENSEN's Übersicht der korrelativen Zusammenhänge hinsichtlich der Intelligenz zwischen Verwandten verschiedener Grade und zwischen nicht-Verwandten. Es wird unterschieden zwischen empirischen Befunden und theoretisch zu erwartenden, d. h. solchen, die einfachen Modellannahmen wie polygener Bedingtheit und intermediärer Genwirkung (vgl. LENZ 1970, S. 231) genügen bzw. gezielte Partnerwahl einbeziehen. Es zeigt sich sowohl empirisch als auch theoretisch eine höhere Korrelation der gemessenen Intelligenz je enger der Verwandtschaftsgrad der Personen ist und vice versa. Wie JENSEN's Formel für V_p deutlich macht, beeinflusst eine Reihe von Faktoren sowohl die "Qualität" indivi-

dueller Genotypen (z. B. Homogamie) als auch die Phänotypen (z. B. Milieu). Nur Phänotypen sind - zumindest was den Menschen betrifft - direkt meßbar, wobei die Ergebnisse mit einem Meßfehler behaftet sind, der aus der Ungenauigkeit des Meßinstruments oder sonstiger mit dem Meßvorgang verbundenen Fakten resultiert.

Tabelle 50: Korrelationskoeffizienten von Intelligenztestresultaten, empirische und theoretische Werte
(hier zitiert nach SCHMID (1977))

Korrelationen zwischen	Anzahl der Untersuchungen	Empirischer Median von r^*	Theoretischer Wert ¹	Theoretischer Wert ²
Nicht verwandte Personen				
Getrennt aufgewachsene Kinder	4	-.01	.00	.00
Pflegeeltern und Kind	3	+.20	.00	.00
Zusammen aufgewachsene Kinder	5	+.24	.00	.00
Kollaterale Verwandtschaft				
Vettern 2. Grades	1	+.16	+ .14	+ .083
Vettern 1. Grades	3	+.26	+ .18	+ .125
Onkel(Tante) und Neffe(Nichte)	1	+.34	+ .31	+ .25
Geschwister, getrennt aufgewachsen	33	+.47	+ .52	+ .50
Geschwister, gemeinsam aufgewachsen	36	+.55	+ .52	+ .50
Zweieiige Zwillinge, verschiedengeschlechtlich	9	+.49	+ .50	+ .50
Zweieiige Zwillinge, gleichgeschlechtlich	11	+.56	+ .54	+ .50
Eineiige Zwillinge, getrennt aufgewachsen	4	+.75	+1.00	+1.00
Eineiige Zwillinge, zusammen aufgewachsen	14	+.87	+1.00	+1.00
Direkte Linie				
Grosseltern und Enkel	3	+.27	+ .31	+ .25
Eltern(als Erwachsene) und Kind	13	+.50	+ .49	+ .50
Eltern(als Kind) und Kind	1	+.56	+ .49	+ .50

* Korrelationen nicht korrigiert

1 angenommen gesteuerte Partnerwahl und partielle Dominanz

2 angenommen zufällige Partnerwahl und lediglich additive Gene, d.h. ein-fachtes mögliches polygenisches Modell

Aus der o. a. Tabelle soll hier zunächst auf die Korrelationen zwischen zusammen aufgewachsenen nicht-verwandten Kindern eingegangen werden. Die Werte be-tragen empirisch $r_e = .24$ und theoretisch $r_t = .00$. Bei der empirischen Korrelation handelt es sich um den Median, den ERLÉNMEYER-KIMLING und JARVIK (1963) aus

5 verschiedenen Untersuchungen errechnet haben. Daß die empirische Korrelation deutlich von Null verschieden ist, wird von den Autoren mit zu kleinen Stichproben-
größen oder solchen Stichproben, die wesentlich von einer Panmixie der Gene abweichen, begründet. In diesem Falle könne die genetische Korrelation zwischen mutmaßlich nicht verwandten Individuen beträchtlich von Null verschieden sein. Die 5 Stichproben rekrutieren sich aus Kindern, die je im selben Waisenhaus oder Pflegeheim vom frühen Alter (early age) an zusammen aufgewachsen sind. Wir kennen nicht die Ausstattung dieser Häuser und die erzieherischen Bedingungen unter denen diese Kinder in ihnen aufgewachsen sind. Untersuchungen von SPITZ (1972) und BOWLBY (1973) belegen jedoch ein-
drucksvoll, welche Deprivationen Heimerziehung auf Kinder haben kann. Obgleich ERLÉNMEYER-KIMLING und JARVIK darauf hinweisen, Untersuchungen an Heimkindern z. B. mit Hirnschäden (mental defectives) nicht in ihre Übersicht aufgenommen zu haben, bleibt die Frage offen, ob ein Vergleich von Kindern, die zusammen z. B. in Internaten aufwachsen, ähnliche oder abweichende Korrelationen liefert. Mit anderen Worten: Es kann kaum unterstellt werden, von JENSEN sei auch nur annähernd das mögliche Spektrum gegebener Umwelten in seine Berechnungen einbezogen worden. Vergleichsuntersuchungen sind deshalb notwendig, weil JENSEN u. a. die Umweltvarianz der Intelligenz aus der Korrelation zusammen aufgewachsener nicht-verwandter Kinder schätzt.

Mit Bezug auf o.a. Tabelle schätzt JENSEN (1969, S. 50f) die Umweltvarianz wie folgt:

"For example, we see that the correlation between identical or monozygotic (MZ) twins reared apart is .75. Since MZ twins develop from a single fertilized ovum and thus have exactly the same genes, any difference between the twins must be due to nongenetic factors. And if they are reared apart in uncorrelated environments, the difference between perfect correlation (1.00) and the obtained correlation (.75) gives an estimate of the proportion of the variance in IQ's attributable to environmental differences: $1.00 - 0.75 = 0.25$. Thus 75 percent of the variance can be said to be due to genetic variation (this is the heritability) and 25 percent to environmental variation. Now let us go to the other extrem and look at unrelated children reared together. They have no genetic inheritance in common, but they are reared in a common environment. Therefore the correlation between such children will reflect the environment. As seen in Table (...) this correlation is 0.24. Thus, the proportion of IQ variance due to environment is 0.24."

Aus den Ausführungen von ERLEMEYER-KIMLING und JARVIK geht nicht hervor, wie die Paarbildung zur Berechnung der Korrelationen für den daraus gewonnenen Median vorgenommen wurden, wir unterstellen aber, daß es sich um Zufallspaare, allenfalls altersgleich oder -ählich, handelt. Letztere Parallelisierung ist deswegen notwendig, weil die Vernachlässigung einer altersbedingten aktuellen Intelligenzhöhe bei der Paarbildung zu unsinnigen Korrelationen führen müßte; denn folgt man der BLOOM'schen Intelligenz-Entwicklungskurve (1971,

S. 67), so ist eine möglichst präzise Altersparallelisierung umso notwendiger, je jünger die Probanden und je größer die in die Untersuchung einbezogene Altersstreuung zum Zeitpunkt der Messung sind. Der Grund für diese Maßnahme ist, daß Intelligenztests in verschiedenem Alter der Probanden nicht nur quantitativ, sondern auch qualitativ Verschiedenes messen oder zumindest messen können. Die Annahme qualitativ verschiedener Intelligenz auf verschiedenen Altersstufen, wie dies z. B. BLOOM tut (1971, S. 3), ist nicht unproblematisch, da kaum je definiert wird, was unter qualitativer Veränderung zu verstehen ist (vgl. LANGEVELD 1970). So gesehen, kann nicht ausgeschlossen werden, daß die von JENSEN verwendeten Daten qualitativ verschiedene Intelligenz - auch im Hinblick auf allgemeine Intelligenz sensu SPEARMAN (g-Faktor) - repräsentieren und somit ein weiteres Argument für eine bessere Deskription der Ausgangsdaten gegeben ist.

10.2 Analyse der Längsschnittdaten hinsichtlich Umwelt-
wirkungen auf die Intelligenzentwicklung=====

Generalisierungen von Forschungsergebnissen sind in der Regel nur dann erlaubt, wenn die Untersuchungen bestimmten Bedingungen genügen. Zu diesen Bedingungen gehört hier z. B. die Wahl der Stichprobe(n). Unter "Normalbedingungen" ist es bei Großversuchen kaum jemals möglich, Stichproben zu bekommen, die den Anforderungen der Stichprobentheorie genügen, da die Mitarbeit der Probanden nur freiwillig erfolgen andererseits aber vom Einverständnis bestimmter Institutionen abhängen kann (z. B. Schulleitung, wenn Untersuchungen während der Unterrichtszeit stattfinden sollen, weil man sonst vielfach den erforderlichen Stichprobenumfang gar nicht zustande brächte); Sozialwissenschaftler müssen daher häufig mit Gelegenheitsstichproben (SELG & BAUER 1971, S. 90) vorlieb nehmen. Wie bereits ausgeführt, ist die Längsschnittstichprobe aus mehreren Gründen nicht repräsentativ für MS, NRW oder BRD, d. h. Ergebnisse sind - für sich genommen - nicht generalisierbar. Wieweit dies auch für die Stichproben gilt, auf die JENSEN Bezug genommen hat, speziell die ZZ- und EZ-Gruppen, kann nur vermutet werden. Immerhin berichtet BLOOM (1964, dt. 1971, S. 81) ein sehr interessantes Ergebnis über getrennt aufgewachsene EZ, die (paarweise) dichotomisiert wurden nach Ähnlichkeit resp. Unähnlichkeit ihrer Erziehungsumgebungen. Für 11 Paare getrennt aber

in ähnlichen Erziehungsumgebungen aufgewachsener EZ ermittelte BLOOM eine IQ-Korrelation von $r_s = .91$, und für 8 Paare in unähnlichen Erziehungsumgebungen nur eine $r_s = .24$. Obschon JENSEN sich in anderen Zusammenhängen auf BLOOM's Buch bezieht, gibt es keinerlei Bemerkung zu diesem Befund.

Bei nicht-verwandten Personen, die also keine gemeinsamen Gene besitzen, reflektiert die Korrelation der IQ's zweier solcher Gruppen die Umweltwirkungen auf die Intelligenzentwicklung. Um unsere im folgenden darzustellenden und zu erläuternden Paarbildungen zu "legitimieren", sei nochmals auf JENSEN's Stichproben aus Waisenhäusern und Pflegeheimen verwiesen. Die Tatsache, in solch einem Haus aufgewachsen zu sein, rechtfertigt allein noch nicht die Annahme gleicher Umweltbedingungen für die Betroffenen, zumindest dann nicht, wenn das Altersintervall der Untersuchungsgruppe groß ist. Es könnte nämlich durchaus sein, daß eine Veränderung der Infrastruktur eines Heimes (Personal, Ausstattung, etc.) auf verschiedene Altersgruppen - aber auch in ihnen - unterschiedliche Wirkungen hatte. Entscheidender wäre u. E., die Umwelten nicht so sehr nach äußeren Gegebenheiten zu definieren, sondern vielmehr nach Kriterien, die die Ausbildung/Entfaltung der Intelligenz maßgeblich beeinflussen. Wir haben deshalb versucht, "experimentelle, in gleicher erzieherischer und familiärer Umgebung aufgewachsene Zwillinge" (expt) zu bilden:

- zunächst wurden die Schüler gemäß ihrer Schichtzugehörigkeit (Beruf des Vaters) den 5 Hauptgruppen (s. S.) zugeordnet, wobei die Sonderschüler/innen außer Betracht blieben. Zur Paarbildung wurde jedoch auch die 2. Ziffer der Berufscodierung mit verwendet.
- innerhalb der 5 Gruppen wurde nach Geschlecht differenziert. Diese Maßnahme schien geboten, da die Diskriminanzanalysen hochsignifikante Intelligenzstrukturunterschiede zwischen Jungen und Mädchen ergeben haben; demnach auch hier eine getrennte Betrachtung angemessen erschien.
- zur weiteren Homogenisierung wurden die Schüler/innen nach der zum Zeitpunkt der Untersuchung 1968 besuchten Schulart (und Klasse) getrennt.
- schließlich wurde noch nach dem Alter (in 2 Halbjahren) unterschieden.

Es waren somit - gemäß den ersten beiden Spiegels-trichen - insgesamt $5 \times 2 = 10$ Gruppen zu bilden. Oberstes Prinzip dabei war es, Paare zu finden, die bezüglich der 2. Ziffer des Berufscode gleich oder möglichst ähnlich waren und dabei derselben Schulart angehörten. Es wurden keine Paare zwischen verschiedenen Schularten gebildet. Sodann wurde versucht, die Paare auch noch hinsichtlich Schulklassenzugehörigkeit und Alter gleich zu halten. Dadurch, daß die Aufteilung nach dem Geschlecht nicht immer zu einer geraden

Anzahl von Schülern innerhalb der drei Schularten (HS, RS, OS) führte und von vier Schülern der Beruf des Vaters nicht bekannt war, mußten einige Schüler ausgeschieden werden. In der Regel war es jeweils ein Schüler mit der höchsten Ziffer der zweistelligen Berufscodierung, weil dies unser Hauptkriterium war und wir immer mit der niedrigsten Codeziffer bei der Paarbildung begannen, so daß sich ein etwa auszuscheidender Schüler fast von selbst ergab.

Die "Güte" der Paarbildung wurde korrelationsstatistisch überprüft, ohne jedoch anschließend irgendwelche Korrekturen vorzunehmen. Berechnet wurden für jede der 10 Paar-Gruppen Rangkorrelationen über

- die 2. Ziffer des Berufscodes
- das Alter (2 Halbjahre)

(Eine Berechnung der Korrelationen zwischen den Schularten war nicht notwendig; sie hätte immer den Wert 1 ergeben, da beide Paarlinge immer der gleichen Schulart angehörten.)

Die eigentlichen Informationen bezüglich der Umweltvarianz sollten anschließend die Rangkorrelationen über die GSW 9 - 10 und GSW 13 - 14 liefern. (Es sei nur der Ordnung halber erwähnt, daß diese Werte bei der Paarbildung unberücksichtigt geblieben sind.)

Vor der Berechnung der GSW-Korrelationen hatten wir in Anlehnung an die empirischen Koeffizienten in

JENSEN's Tabelle folgende Hypothesen formuliert:

- 1) Obgleich die Paarlinge erbverschieden und die einzelnen Gruppen sehr stark homogen sind, erwarten wir dennoch positive Korrelationen.
- 2) Die Korrelationen werden zwischen den Geschlechtern nicht wesentlich verschieden ausfallen, da wir GSW (durch Transformation in IQ zu verwandeln) verwendet haben und die Geschlechter sich darin insgesamt nicht unterscheiden.
- 3) Obgleich infolge der Homogenisierung vor allem über den Sozialstatus nicht auszuschließen ist, daß die Varianzkomponente V_{AM} (:= gezielte Partnerwahl) hier einen Einfluß hat, liegen die mittleren Korrelationen deutlich unter denen von getrennt aufgewachsenen Geschwistern, für die JENSEN einen Wert von $r_e = .47$ angibt.

Die Ergebnisse zeigt Tabelle 51. Da in der Berufsgruppe I, Mädchen, nur ein Schülerpaar gebildet werden konnte, mußte dies unberücksichtigt bleiben. Außerdem ist die Anzahl der Paare in den Gruppen I und V jeweils nicht hinreichend groß, so daß die hieraus sich ergebenden Resultate nur in Verbindung mit den Gruppen II, III und IV zu bewerten sind, in Sonderheit, wenn sie eine bei letzteren Gruppen vorfindbare Tendenz stützen sollten. Trotz der teilweise geringen Gruppen-

größen liegen die Korrelationen über die Väterberufe in der Größenordnung gut beieinander. Dabei muß man sich jedoch daran erinnern, daß $r_s = 0.00$ in den Gruppen V_m und V_w darauf zurückzuführen ist, daß hier die Kovarianz Null ist (es gibt nur die Codeziffer 5). Die ziemlich stark streuenden Korrelationen über das Alter könnten zurückzuführen sein auf den höheren Anteil jüngerer Schüler an der Gesamtstichprobe und die unterschiedliche Verteilung auf die 10 Gruppen.

Tabelle 51: Korrelationen von Schülergruppen, die nach dem Beruf des Vaters, dem Geschlecht, dem Alter und der Schulartzugehörigkeit parallelisiert wurden (ohne Sonderschüler). Tests: GSW 8 - 10 und GSW 13 - 14

Berufsgruppe (I2)		I	II	III	IV	V	$\bar{r}_s$	Median
Beruf d. V. 2. Ziffer	m	.50	.88	.73	.60	.80	.30	.80
	w	-	.47	.60	.72	.00	.48	.53
Alter (2Halbjahre)	m	.00	.19	1.00	.45	.50	.69	.45
	w	-	.45	.65	.77	.67	.65	.66
GSW 9 - 10	m	.50	.36	.31	.39	.50	.42	.39
	w	-	.34	.57	.48	1.00	.81	.54
$\bar{r}_s$ über z'		.50	.35	.45	.44	.94	.63	.50
GSW 13 - 14	m	-.50	.10	-.11	.46	.40	.07	.10
	w	-	.46	.28	.49	.75	.53	.47
$\bar{r}_s$ über z'		-.50	.29	.18	.47	.60	.32	.33
Anzahl d. Paare	m	3	9	16	21	3		
	w	1	8	14	18	4		

Kommen wir nun zu den Befunden, die im Zusammenhang mit den Hypothesen stehen:

- (1) Für GSW 9 - 10 sind alle Korrelationen positiv; mit einer Streubreite von .31 - 1.00. Die Maximumkorrelation bei Vw muß im Zusammenhang mit der geringen Anzahl der Paare ($N=4$) gesehen werden. Auch für GSW 13 - 14 sind 7 von 9 Koeffizienten positiv (Streubreite -.50 bis .75). Hier ist $r_g = -.50$ bei Im ebenfalls auf die geringe Gruppengröße zurückzuführen. Unklarheit besteht über die negative Korrelation $r_g = -.11$ bei IIm.
- (2) Bis auf einen Koeffizienten (bei IIm) sind alle Korrelationen bei den Mädchen über GSW 8 - 10 und GSW 13 - 14 zum Teil beträchtlich höher als die der Jungen. Dies zeigen besonders die Mittelwerte ($\bar{r}_g$ und Median) der Tabelle 51 und gilt sowohl für GSW 9 - 10 als auch GSW 13 - 14. Die Hypothese 2 hat sich nicht bestätigt.
- (3) Auch Hypothese 3 konnte nicht bestätigt werden, denn von den 18 Koeffizienten sind 7 größer als der Wert (.47), unter dem alle Korrelationen liegen sollten; drei weitere liegen mit .40 bis .46 nur wenig unterhalb jenes Wertes. Aber auch die vier Koeffizienten zwischen .31 bis .39 hatten wir in dieser Höhe nicht erwartet. Insgesamt entsprechen also 14 von 18 Werten nicht der Hypothese. Die zeilen- und spaltenweise berechneten $\bar{r}_g$ (über z') verdeutlichen den Gesamtsachverhalt nochmals.

Wie können nun diese Ergebnisse interpretiert werden? Gehen wir - wie JENSEN - von den Medianen aus, so ergeben sich je nach dem Alter der Schüler verschiedene Erb- und Umweltvarianzen. Für die ca. 10jährigen Schüler ergäbe sich nach JENSEN's Formel eine Erbvarianz von $1 - .50 = .50$, also 50%. Im Alter von ca. 14 Jahren wäre die Erbvarianz für dieselben Schüler auf $1 - .32 = .68$, also 68% angestiegen. Auf diese unterschiedlichen Werte werden wir weiter unten eingehen. Zunächst soll die weiter oben offen gebliebene Frage nochmals aufgenommen werden; nämlich, ob es angemessen ist, zur Bestimmung von Umweltvarianzanteilen Korrelationskoeffizienten zu verwenden. Nach HAYS (1974, S. 535) müssen derartige Koeffizienten "intra-class-correlations" sein, die aus Zufallsstichproben gewonnen wurden. JENSEN geht auf diese Problematik nicht ein. Die Daten von ERLÉNMEYER-KIMLING und JARVIK sowohl über Zwillinge als auch nicht-Verwandten lassen jedoch nicht den Schluß JENSEN's zu, es handele sich bei seinen Schätzungen um gute Annäherungen an die "wahren" Werte für die weiße Bevölkerung der USA. Zumindest der Korrelationsmedian aus den Waisenhaus- und Pflegehauskindern dürfte kaum repräsentativ für nicht-verwandte, zusammen aufgewachsene weiße Kinder der USA sein, so daß hier zur Schätzung der Umweltvarianz sinnvollerweise auch nur das Quadrat des Korrelationskoeffizienten verwendet werden kann, also $r^2 = .24^2 = .58$, entsprechend 5,8% Umweltvarianz. Bezogen auf unsere Daten heißt das,

$1 - .50^2 = .75$, resp. 75% Erbvarianz im Alter von ca. 10 Jahren und $1 - .33^2 = .89$, resp. 89% Erbvarianz im Alter von ca. 14 Jahren.

Eine für beide Geschlechter getrennte Berechnung ergibt:

Mädchen, ca. 10 Jahre: Erbvarianz 71%

Jungen, ca. 10 Jahre: Erbvarianz 85%

Mädchen, ca. 14 Jahre: Erbvarianz 78%

Jungen, ca. 14 Jahre: Erbvarianz 99%

Die Differenz zwischen den einzelnen Erbvarianzen und 100 entspräche jeweils der Umweltvarianz. Da unsere Korrelationen infolge der Paarbildung einen Varianzanteil haben können, der a conto gezielter Partnerwahl gehen könnte, müßte dieser Teil von der Erbvarianz subtrahiert werden (ca. 5%). Um Mißverständnissen vorzubeugen; diese Überlegungen gelten nur unter der Annahme der JENSEN-Formel.

In einem Abschnitt erörtert JENSEN (1969, S. 52f) die Frage direkter Messung der Umweltvarianz. Dies sei einfach durch Korrelation der IQ's der Probanden mit den Umweltratings zu erreichen, aber legitimerweise nur mit Adoptivkindern möglich und wenn sichergestellt sei, daß keine gezielte Zuweisung der Kinder durch die Adoptionsvermittlung erfolgt sei, da sonst ein Teil der so erhaltenen Varianz genetisch bedingt sei (Homo-gamie-Effekt). Zur Demonstration zitiert JENSEN zwei

Arbeiten, LEAHY (1935) und BURKS (1928). LEAHY berichtet eine Korrelation zwischen IQ's/mehrere Umweltindices von $r = .20$, was einer Umweltvarianz von 4% ($r^2 \times 100$) entspräche, während BURKS eine multiple Korrelation (korrigiert wegen mangelnder Reliabilität der Meßinstrumente) von $r = .42$, resp. 18% Umweltvarianz erzielte. LEAHY's Wert von 4% ist nach JENSEN deswegen zu niedrig ausgefallen, weil die Umweltindices nicht genügend "feinkörnig" (fine-grained) waren und kein multipler Korrelationskoeffizient berechnet wurde, der üblicherweise beträchtlich höher sei. Diese Nachteile weist BURKS' Arbeit nicht auf, so daß JENSEN dieses Resultat als gute Annäherung an seinen Mittelwert (20%) interpretiert. Weniger die Koeffizienten selbst, sondern die Tatsache, daß JENSEN es hier offenbar für richtig hält, r^2 zur Varianzschätzung zu verwenden, verwundert einigermaßen, denn es besteht stichprobentheoretisch kein prinzipieller Unterschied zwischen einerseits LEAHY's und BURKS Untersuchungen und andererseits den Waisenhaus- und Pflegehauskindern ERLÉNMEYER-KIMLING und JARVIK's. Auch unsere Umweltkriterien scheinen noch recht grob für eine präzise Varianzschätzung. Es kann daher erwartet werden, daß die Erziehungsbedingungen genauer berücksichtigende Stichproben größere Umweltvarianzen zutage bringen werden. Immerhin liegen unsere Werte jetzt schon zum Teil höher als BURKS-Ergebnisse. Ferner stimmt unser Median (.33) gut mit einem Befund JENSEN's überein: JENSEN (1969a) fand zwischen nicht-verwandten,

aber unter gleichen SÖS-Bedingungen aufgewachsenen Personen eine IQ-Korrelation von $r = .30$, entsprechend einer Varianz von ca. 9%. JENSEN läßt dieses offensichtlich zu niedrige Resultat allerdings außer Betracht!

Von nicht geringer Bedeutung könnte eine genauere Untersuchung der von uns gefundenen unterschiedlichen Koeffizienten zwischen den Geschlechtern sein. Auch dieses Problem ist von JENSEN und anderen nicht hinreichend aufgenommen worden. JENSEN (1969a) gibt in seiner Replik auf KAGAN's Kritik eine Tabelle von SHIELDS (1962) wieder, die Korrelationen von zusammen und getrennt aufgewachsenen EZ über 4 Intelligenz- und Fähigkeitstests, der Körpergröße, dem Gewicht, Extraversion und Neurotizismus enthält. Eine für beide Geschlechter getrennte Angabe von Korrelationskoeffizienten erfolgt nur für Körpergröße und Gewicht. Es mutet merkwürdig an, daß den psychologischen Merkmalen keine geschlechtsspezifische Aufmerksamkeit JENSEN's zuteil geworden ist.

Die Analysen liefern noch einen weiteren, u.E. wichtigen Befund, der als Faktum von JENSEN zwar nicht übersehen wurde, aber doch nicht die nötige Beachtung hinsichtlich der Konsequenzen gefunden hat: JENSEN weist in einem Abschnitt darauf hin, die Erbllichkeit der Intelligenz sei keine Konstante, sondern Schätzungen der Erbllichkeit seien spezifisch für die zugrunde liegende Population und abhängig vom Zeitpunkt, den Umständen der

Messung sowie den verwendeten Messinstrumenten. Und er stellt fest, die in der Literatur mitgeteilten Werte der Erbllichkeit seien weder für jegliche Umwelten verschiedener Populationen noch für ein und dieselbe Population zu verschiedenen Zeitpunkten gültig. Diese schon in der Formel zur Berechnung V_p explizit enthaltene Bedingtheit, wird noch präzisiert durch die Formulierung, es sei ein Faktum, daß die Erbllichkeit der Intelligenz mit zunehmenden Alter steige. Genau diese Altersabhängigkeit der Erbllichkeit wird durch unsere Daten bestätigt. Bei der ersten IQ-Messung waren unsere Schüler schon ca. 10 Jahre alt. Es darf somit angenommen werden, daß die Erbllichkeit noch erheblich abnimmt, wenn etwa bei einer neuen Untersuchung auch noch frühere Messungen durchgeführt werden. Den Differenzbetrag zwischen Messungen im Alter von 6 und 10 Jahren würden wir - mit Blick auf BLOOM's nicht-lineare Entwicklungskurve des IQ - höher ansetzen als zwischen 10 - 14 Jahren. Bezogen auf unsere Daten hieße das, die Erbllichkeit des IQ liegt unter 60% bei 6jährigen. Eine weitere Extrapolation hätte natürlich eine weitere Reduktion der Erbllichkeit zur Folge.

Man kann also innerhalb des positiven Spektrums (0 bis 1) in einer Längsschnittstudie je nach Alter der Probanden zum Zeitpunkt der Messung beinahe jeden beliebigen Wert erzielen. Sind vielleicht so ein Teil der Streubreiten der Korrelationskoeffizienten ERLLENMEYER-KIMLING und

JARVIK's zu erklären? JENSEN verfolgt seinen eigenen Gedanken jedoch nicht weiter, gibt aber auch nicht an, auf welches Alter sich seine mittlere Schätzung (80%) beziehen soll! Unsinnig wäre es allerdings, einen Mittelwert (über mehrere Altersjahrgänge) zu bilden und darauf etwa eine Argumentation für oder gegen bestimmte Förderprogramme zu stützen. Mittelwerte dieser Art wären darüber hinaus auch völlig nutzlos zur Konstruktion und Durchführung gezielter pädagogisch-psychologischer Maßnahmen zur individuellen Intelligenzförderung. U. E. wäre es allerdings auch falsch zu folgern, die Frage der Erbllichkeit der Intelligenz könne oder dürfe künftig vernachlässigt werden. Davon könnte allenfalls gesprochen werden, wenn sich in entsprechenden Untersuchungen über

- alle Altersgruppen aller Sozialschichten
- weder geschlechts- noch schichtspezifisch

bedeutsame Unterschiede ergeben sollten.

(Zudem müßte eine Überprüfung der Wirkung spezifischer Varianzkomponenten (z. B. Homogamie) für eine bestimmte Population zum Ergebnis haben, auch eine völlige Vernachlässigung von genetischen Faktoren in der politisch-pädagogischen Erziehungswirklichkeit werde langfristig keinerlei Effekte haben, die den Postulaten des Grundgesetzes der freien Entfaltung der Persönlichkeit und maximaler Gerechtigkeit zuwider laufen. Es wäre töricht und verantwortungslos, die Frage (a) ohne genaue Prüfung ad acta zu legen oder (b) im Falle einer negativen Antwort nicht die notwendigen Konsequenzen zu ziehen. Es sollte nicht übersehen werden, daß auch von hier eine

Um-/Neuorientierung der Werte in der Gesellschaft ausgehen kann.)

Falls sich dies jedoch nicht bestätigt, könnten die Ergebnisse Hinweise liefern, wann gezielte Förderprogramme mit größter Effizienz für alle Beteiligten eingesetzt werden bzw. müßten. Die Fülle bisher vorliegender Befunde, die für eine möglichst frühe - auch spielerische - Förderung sprechen, können von allen Sozialschichten für sich reklamiert werden. Sollte es sich aber zeigen, daß z.B. die Erbvarianz der Intelligenz bei Unterschichtkindern schneller steigt als bei Mittel- und Oberschichtkindern und sich eher als erwartet auf einem Höchstwert stabilisiert, so wäre die bereits erwähnte Forderung des DEUTSCHEN BILDUNGSRATS im Strukturplan (1970) nach Chancengleichheit erneut zu stellen.

Bisher war nur von Korrelation die Rede und nicht von IQ-Differenzen und möglichen Erklärungen dafür. Weil unsere Gruppen besonders der Schichten I und V zu klein sind, soll darauf bewußt verzichtet werden. Die Tabelle 52 soll jedoch einen Überblick geben über die Mittelwerte der miteinander korrelierten Paarreihen (mit A und B gekennzeichnet) sowie die GSW der Paare, die jeweils die größte GSW-Differenz aufgewiesen haben.

Tabelle 52: Arithmetische Mittelwerte und jeweils größte absolute Differenzen der jeweils korrelierten Paarreihen A und B, getrennt nach SÖS und Geschlecht, für GSW 8 - 10 und GSW 13 - 14
 Normmittelwert = 50
 A_i und B_i bilden jeweils 1 Paar $i = 1, 2 \dots n$

Tests		Berufsgruppen (SÖS) und Geschlecht									
		I		II		III		IV		V	
		m	w	m	w	m	w	m	w	m	w
W 8-10	Reihe A	62.0	55.8	52.7	53.5	49.5	50.6	47.2	48.0	43.2	
	Reihe B	54.0	50.3	54.5	56.6	52.1	50.3	50.2	46.7	45.0	
Paar mit größter Differenz, absolut		55/53	58/47	60/44	51/77	39/57	65/39	54/39	47/41	50/57	
		12	11	16	26	18	26	15	6	7	
W 13-14	Reihe A	58.3	54.4	52.1	53.4	50.3	52.9	47.3	47.3	42.5	
	Reihe B	52.7	53.4	52.0	54.5	51.4	50.0	47.4	44.3	44.7	
Paar mit größter Differenz		66/48	48/62	51/40	66/44	66/47	66/40	34/52	49/43	38/47	
Differenz, absolut		18	14	11	22	19	26	18	6	9	

Die Mittelwerte liegen für die jeweiligen Paarreihen A und B wider Erwarten gut beieinander, abgesehen von Gruppe Im, was durch die geringe Anzahl der Probanden erklärt werden kann. Die größten Differenzen innerhalb von Paaren der "experimentellen Zwillinge" sind dagegen sehr beträchtlich. Sie können als Indikator dafür gewertet werden, daß "feinkörnigere" Umweltindizes die gefundenen Korrelationskoeffizienten noch erhöhen werden, was gleichbedeutend mit einer Steigerung der Umweltvarianz wäre.

Abschließend wollen wir nochmals an BLOOM anschließen, der ja eine Homogenisierung der Zwillinge von NEWMAN et al. (1937) nach der Erziehungsumgebung vorgenommen hat, und selbst einige zusätzliche Berechnungen durchführen, die noch andere Homogenisierungsaspekte betonen.

10.3

EXKURS:

- a) Neue Analyse einiger Daten der Untersuchung von NEWMAN, FREEMAN und HOLZINGER (NFH) 1937 ⁺

In ihrer o.a. Arbeit berichten NFH unter anderem Forschungsergebnisse von 19 Paaren getrennt aufgewachsener eineiiger Zwillinge (EZg). Da die Autoren eine Fülle von Einzeldaten der Zwillinge mitgeteilt haben, ist es möglich, einige Berechnungen und Gruppierungen erneut durchzuführen.

- a) Korrelationskoeffizienten:

Die Autoren teilen für die EZg bezüglich der Intelligenz (Binet-IQ) einen Korrelationskoeffizienten von $r=0,67$ mit (NFH 1937, S. 347). Dieser Koeffizient beruht auf der Formel von PEARSON. An deren Verwendung sind jedoch eine Reihe von Bedingungen geknüpft, wie Stichprobenumfang und -auswahl, um nur einige zu nennen (vgl. SIEGEL 1956, KREYSZIG ³1968). Da diese

⁺) Die für die neuen Analysen notwendigen Berechnungen wurden von uns durchgeführt.

Bedingungen u.E. in der Untersuchung von NFH kaum als erfüllt anzusehen sind, geben wir der Rangkorrelation r_s nach SPEARMAN den Vorzug. Dabei entsteht jedoch ein Problem: Die Aufteilung der Paare auf die beiden Rangreihen, d.h., gesucht ist ein Kriterium, nach dem jeweils die beiden Partner eines Zwillingspaares entweder der Rangreihe 1 oder 2 zugeordnet werden. Eine Möglichkeit hierzu ist das Randomverfahren. Eine andere ist u.E. die Doppelzuordnung nach folgendem Schema:

Zwillingspaar	IQ-Werte	
	Reihe 1	Reihe 2
I	x_1	x_2
	x_2	x_1
	x_3	x_4
II	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
m	x_{n-1}	x_n
	x_n	x_{n-1}

Da wir jedoch NFH's Korrelationen nach PEARSON mit Rangkorrelationen vergleichen wollen, haben wir einmal die Anordnung so beibehalten, wie sie von NFH in deren Tabelle 87 angegeben wurde. Zum Vergleich dazu haben wir dann noch Rangkorrelationen nach dem Schema der Doppelzuordnung berechnet. Tabelle 53 zeigt eine Gegenüberstellung der Ergebnisse.

Tabelle 53 : Korrelationen verschiedener Gruppen von EZg, berechnet für den Binet-IQ der Untersuchung von NEWMAN, FREEMAN u. HOLZINGER (= NFH) N = Anzahl

Gruppe		N- Paare	NFH- Ordnung r_s	N- Paare	Doppelzu- ordnung r_s
Gesamt-N		19	0,55	38	0,60
Alter	Jüngere	5	0,80	10	0,76
	Ältere	14	0,41	28	0,43
Geschlecht	männlich	7	0,82	14	0,76
	weiblich	12	0,57	24	0,59
Erziehungs- umgebung n. BLOOM	ähnlich	11	0,91	nicht berechnet, da die Grenze nicht ganz klar ist.	
	unähnlich	8	0,24		

Die Koeffizienten liegen für beide Berechnungsmodi gruppenspezifisch dicht beieinander, für das Gesamt-N jedoch - wie erwartet - unter dem nach PEARSON errechneten Wert (0,67).

Interessanter als dieses Resultat sind die Ergebnisse bezüglich Alter, Geschlecht und Erziehungsvorteil resp. -nachteil, von denen die Differenzierung innerhalb der beiden erstgenannten Kriterien u.W. bisher noch nicht erfolgt ist, wenngleich die Notwendigkeit hinsichtlich des Alters bereits gefordert wurde (vgl. KAMIN 1974, S. 56).

b) Alter der NFH-Zwillinge:

Es ergeben sich mit $r_g=0,76$ für die Jüngeren und $r_g=0,43$ für die Älteren deutliche Unterschiede hinsichtlich der Ähnlichkeit der Intelligenz unter den Zwillingspaaren, die in dieser Form an sich nicht auftreten dürften, da ja der IQ (bzw. die Standardisierung) den Alterseinfluß eliminiert. Die Unterschiede der Korrelationskoeffizienten sind statistisch jedoch nicht signifikant. Gründe für die Unterschiede können in der Testkonstruktion und den damit zusammenhängenden Problemen aber auch gerade bei den Älteren in der vermutlich weiter auseinanderklaffenden Vertrautheit mit der Situation der Testdurchführung, die ja nicht ohne Einfluß auf die Ergebnisse ist (vgl. KAMIN 1974), liegen.

c) Geschlecht der NFH-Zwilligen:

Die Berechnung der IQ-Korrelation - getrennt für die Geschlechter - ergeben folgende Werte: $r_g=0,76$ (männlich) und $r_g=0,59$ (weiblich). Die Unterschiede sind zwar deutlich aber statistisch nicht signifikant.

d) Erziehungsumgebung der NFH-Zwillinge:

Hierzu werden in der Tabelle die Koeffizienten wiedergegeben, die von BLOOM (1964, dt. 1971, S. 81) berechnet wurden.

- b) Neue Analyse einiger Daten der Zwillingsuntersuchung von SHIELDS (1962).

SHIELDS Zwillingsgruppen "S" (separated) und "C" (Control) sind ebenfalls groß genug bzw. noch größer als die NFH's, um sie nach dem Geschlecht zu trennen und für männliche und weibliche Probanden getrennte Korrelationen zu berechnen. Tabelle 54 zeigt die Ergebnisse.

Tabelle 54 : Korrelationen der Zwillinge von SHIELDS, zusammen und nach Geschlechtern getrennt

	(A) Domino	Tests (B) Mill-Hill	IQ A + (2B)
<u>Gruppe "S"</u>			
Gesamt-N=37	.77	.78	.79
Jungen =13	.62	.71	.67
Mädchen =24	.81	.82	.83
<u>Gruppe "C"</u>			
Gesamt-N=67	.74	.76	.76
Jungen =21	.89	.62	.68
Mädchen =45	.64	.78	.74

Auch diese Resultate zeigen - wie bei NFH - deutliche Unterschiede in den Korrelationskoeffizienten. Die Höhe der Koeffizienten ist nicht gleichsinnig. Während

bei NFH beide Korrelationen der männlichen Pbn höher waren, haben bei SHIELDS die weiblichen Pbn in drei von vier Fällen die höheren Werte.

Insgesamt sollten diese Ergebnisse dazu veranlassen, bei zukünftigen Zwillingsvergleichen möglichst homogene Gruppen zu bilden, da andernfalls zu viele Informationen verloren gehen und kontroverse Auffassungen vielleicht entdramatisiert werden können.

TEIL II

DIE KONTROLLUNTERSUCHUNG

11.0 ÜBEREINSTIMMUNGSVALIDITÄT VON INTELLIGENZTESTS

11.1 Der Stellenwert der Kontrolluntersuchung

In Teil I haben wir dargestellt, daß die Längsschnitt-Untersuchungsstichprobe auch gleichzeitig zur Standardisierung der dort verwendeten Testbatterien und Einzelbegabungstests herangezogen wurde. Darüberhinaus sind die Tests auch als Diagnostika in der Untersuchung selbst verwendet worden. Nun wird jedoch aus testkonstruktionstheoretischen Gründen gefordert, die Validitätsprüfung von Tests nicht auch an der Stichprobe vorzunehmen, an der auch die Aufgabenanalyse erfolgt ist, sondern hierzu eine andere Stichprobe (Kreuzvalidierung) zu verwenden, um mögliche systematische Fehler aufzudecken (LIENERT, 1967, S. 53).

Da es sich bei der Hauptuntersuchung jedoch um eine Längsschnittanalyse des Schulerfolgs oder -mißerfolgs von Schülerinnen und Schülern seit Beginn ihrer Einschulung handelt, der Stichprobe, die bereits 1960/61 mit damals verfügbaren Schulreifetests untersucht worden war, für deren Jahre

später wiederholte Untersuchungen keine geeigneten Tests vorhanden waren, so daß die später verwendeten Tests und Fragebögen entweder aus dem amerikanischen Übersetzt oder selbst konstruiert und standardisiert werden mußten (KEMMLER 1967, KEMMLER & LANGHEINRICH 1967, KEMMLER & LANGHEINRICH 1969, KEMMLER 1976), schien es wünschenswert, die Validität der PMA 13-14 im Vergleich mit geeichten Tests (Batterie), die ebenfalls auf der Faktorentheorie der Intelligenz von THURSTONE (1938, 1954) basieren, zu ermitteln. Dies ist bei dem Prüfungssystem für Schul- und Bildungsberatung (PSB) von HORN (1969) und dem Frankfurter Analogietest (FAT 7-8) von BELSER, ANGER und BARGMANN (1968) gegeben, so daß diese Tests zur Validitätsprüfung herangezogen wurden.

Das Hauptziel der Kontrolluntersuchung ist also die Ermittlung der Übereinstimmungsvalidität verschiedener Testserien. Dieses Vorhaben sollte sich jedoch nicht auf die Berechnung von Korrelationsmatrizen und Faktorenanalysen beschränken. Gleichzeitig sollte geprüft werden, ob sich (inhaltliche) Ergebnisse der Analyse der Längsschnittuntersuchung reproduzieren ließen.

11.2 Die Kontrollstichprobe

Aus mehreren Gründen kann diese Validitätsprüfung nur als ein Schritt für eine umfassende Prüfung verstanden werden, denn die Zusammensetzung der Kontrollstichprobe sollte z.B. möglichst ein Abbild der Längsschnittstichprobe sein, um in einem Rückschlußverfahren die Ergebnisse der Längsschnittuntersuchung noch einmal bewerten zu können.

Die Überlegung dazu ist folgende:

Wenn

- (a) die zu vergleichenden Tests zufriedenstellende Testkoeffizienten aufweisen,
- (b) sich in der Kontrollstichprobe etwa bei Untergruppenvergleichen - getrennt nach den Subtests der verwendeten Testbatterien - Übereinstimmende Resultate ergeben,

so würde dies

- (c) gleichlautende Resultate der Längsschnittanalyse stützen, in dem Sinne, daß im Maße der Übereinstimmung der Ergebnisse mit der Kontrollstichprobe die Tatsache der Standardisierung der PMA 13-14 an der Längsschnittstichprobe vermutlich keine systematischen Fehler enthält. Hierauf wird im Einzelfall noch zurückzukommen sein.

Derartige Ergebnisübereinstimmungen wären jedoch nicht so ohne weiteres etwa für die vergleichbaren

Schüler Münsters oder gar überregional repräsentativ. Diese Einschränkung des Gültigkeitsbereichs folgt aus der Organisation der Stichproben. Mit der Stichprobenfrage ist immer auch die Frage der Repräsentativität der zu erwartenden Forschungsergebnisse gestellt. Es werden zwei Arten von Stichproben unterschieden:

- (1) Zufallsstichproben (einschl. Spezifikationen),
- (2) Quotenstichproben.

Bei der Frage, welche Stichprobenart für die zu bearbeitende Problematik die angemessenste sei, stößt man zumindest bei Fragen aus der pädagogischen Psychologie - wenn eine größere Probandenzahl erforderlich ist - schnell auf äußere Gegebenheiten, die meistens kaum völlig zu umgehen sind: So sind z.B., die Probanden (in der Regel Schüler) in größerer Zahl am ehesten dann für eine Mitarbeit zu gewinnen, wenn das Vorhaben während der regulären Unterrichtszeit stattfindet. Dies wiederum hat meist zur Folge, daß man eine Quotenstichprobe wählt. In vielen Fällen dürfte es aber kaum möglich sein, die Voraussetzungen, die an eine Quotenstichprobe gestellt werden, hinreichend zu erfüllen (SELG & BAUER 1971, S. 89ff), denn wenn man schließlich Schulen "gefunden" hat, die zur Mitarbeit bereit sind, so wird häufig von seiten der Schule der verständliche Wunsch geäußert, möglichst immer ganze Klassen in den Versuch einzubeziehen, damit nicht etwa eine Gruppe normalen Unterricht erhalte, der von den Probanden nicht nachzuholen sei.

Ein juristisches Problem ergibt sich ferner aus der Aufsichtspflicht der Schulleitung über die Schüler. Sie macht eine Zusammenführung von Schülern während der Unterrichtszeit außerhalb der Schule, sofern nicht eine Genehmigung der Schulaufsichtsbehörde vorliegt, für den Untersucher aus ver-sicherungsrechtlichen und für die Schulleitung darüber hinaus vermutlich noch aus disziplinarrecht-lichen Gründen schwierig. In den meisten Fällen wird der Untersucher jedoch schon aus ökonomischen Überlegungen oder auch nur finanziellen Gegeben-heiten akzeptieren müssen, sein Vorhaben in dem ihm möglichen Rahmen durchzuführen. Für experimen-telle Forschungspläne, etwa die Prüfung verschiede-ner Unterrichtsmethoden, besteht dann noch ein zusätzliches Problem, nämlich das der Intraklassen-korrelation (ECKEL 1969, 1970; KLAUER 1970, 1973).

Für die Zusammenstellung der Kontrollstichprobe war folgendes zu beachten:

- Es sollten etwa im Verhältnis 2 : 1 : 1 Hauptschü-ler, Realschüler und Oberschüler (Gymnasiasten) in der Stichprobe vertreten sein,
- Die HS-, RS- und OS-Gruppen sollten beide Geschlech-ter etwa im Verhältnis 1 : 1 enthalten,
- Die Schüler sollten zum Zeitpunkt der Untersuchung zwischen 13;6 - 14;5 Jahre alt sein. Als Stichtag wurde der 15. 10. 1970 festgelegt.

Diese Bedingungen konnten realisiert werden, ohne den "Schulbetrieb" jeweils nennenswert zu stören. Die Berücksichtigung weiterer Klassifikationsmerkmale wie "Beruf des Vaters", "Einschulungstermin" oder "Versetzungen" war im vorhinein wegen zu großer Inanspruchnahme der Schulorganisation nicht möglich, so daß beschlossen wurde, diese Angaben von den Schülern selbst zu erfragen. Da aber bei einer RS-Klasse die Frage nach den Versetzungen versehentlich nicht ausdrücklich gestellt wurde und dadurch, daß auch eine nicht geringe Anzahl von Hauptschülern (24% von 58, insgesamt 29 Schüler, 25% von 116) hierzu keine Angaben gemacht hat, bleibt dieses Merkmal bei der Analyse weitgehend unberücksichtigt. Noch lückenhafter war die Beantwortung der übrigen Fragen, so daß hier auch auf eine Darstellung verzichtet wird. Tabelle 55 zeigt die genauere Zusammensetzung der Kontrollstichprobe.

Tabelle 55 : Zusammensetzung der Kontrollstichprobe ⁺⁾
nach Schulart, Klasse, Versetzungen, Geschlecht und Alter (in Halbjahren)

Schul- art	Klasse	Ver setzung	Jungen		Mädchen		
			Ältere	Jüngere	Ältere	Jüngere	
HS	8	n.v.	1	1	1	2	5
		i.v.	4	8	4	8	24
		k.A.	0	0	0	0	0
HS	9	n.v.			1		1
		i.v.	6	1	4	3	14
		k.A.	5	2	5	2	14
RS	8	n.v.	3	1			4
		i.v.	1	10			11
		k.A.	0	0	6	9	15
OS	8	n.v.	1	4	0	3	8
		i.v.	5	4	3	8	20
		k.A.	0	0	0	0	0
Summe			26	31	24	35	116

Es bedeuten: n.v. := mindestens einmal nicht versetzt
 i.v. := immer versetzt
 k.A. := keine Angaben

⁺⁾ Den Schulleitern, Lehrern und Schülern sei an dieser Stelle für ihre Mitarbeit nochmals herzlich gedankt.

Eine Zusammenfassung der Probanden nach Schulart (HS, RS, OS), Geschlecht und Alter (Jüngere/Ältere) zeigt im Sinne statistischer Prüfung nur erwartungsgemäß zufällige Gleichverteilungsabweichungen. Eine Ausnahme ergibt die Gleichverteilungsprüfung der HS nach Klasse (8./9.) und Alter.

Hauptschüler	Ältere	Jüngere	
8. Klasse	10	19	29
9. Klasse	21	8	29
Summe	31	27	58

Die Verteilungsprüfung ist sehr signifikant, d.h. wie zu erwarten, befinden sich unter den Schülern der 8. Klassen wesentlich mehr jüngere, denn diese Schüler sind ja im Durchschnitt ein Jahr jünger als die der 9. Klassen. Es bleibt hier der Grund für die Aufnahme von Schülern der 9. Klassen in unsere Kontrolluntersuchung nachzutragen: Die in ihrer Schullaufbahn am weitesten fortgeschrittenen Schüler der Längsschnittuntersuchung besuchten zur Zeit der Testerhebung 1968 die 9. Klasse. Dies galt es hier zu berücksichtigen. Allerdings schien uns dies angesichts der geplanten Größe der Kontrollstichprobe und deren Untergruppen nur bei den Hauptschülern sinnvoll.

11.3 Die verwendeten Tests, Durchführung und
Auswertung der Kontrolluntersuchung====

Beim "Prüfsystem für Schul- und Bildungsberatung" (P-S-B) von HORN (1969) handelt es sich um eine aus 10 Subtests bestehende Testbatterie. Es wurden aus dem "Leistungsprüfsystem" (L-P-S) von HORN (1962) jene Teiltests ausgewählt, "die für schulische Zwecke besonders wichtig sind" ... und "von großer diagnostischer Relevanz" (HORN 1969, S. 4). Der Autor hat das L-P-S und damit implizit auch das P-S-B in Anlehnung an die Primärfaktorentheorie von THURSTONE konstruiert. Die Tests 1 + 2 (Allgemeinbildung) und Test 5 (Worteinfall) sowie Test 6 (Ratefähigkeit) entsprechen THURSTONE's "Verbal-Factor", die Tests 3 + 4 (Denkfähigkeit) dem "Reasoning-Factor" und der Test 7 (Raumvorstellung) dem "Space-Factor"; die Tests 7 + 8 (Gliederungsfähigkeit) messen nach Horn zusammen "Technische Begabung". THURSTONE's "Number-Factor" wird vom Test 9 (Addieren) und der "Perceptual Speed-Factor" vom Test 10 (Wahrnehmungstempo) gemessen.

Auch der "Frankfurter Analogietest" (FAT 7-8) von BELSER et al. (1965) basiert auf THURSTONE's Theorie der Primärfähigkeiten. Der FAT 7-8 "erfaßt als Intelligenztest im wesentlichen den Thurstone'schen Faktor R (Reasoning), d.h. die Fähigkeit, Gesetzmäßigkeiten

zu erkennen und anzuwenden" (BELSER et al. 1965, S. 3). Der FAT 7-8 besteht aus einem einzigen Test mit 84 Aufgaben mit Mehrfachwahlantworten.

Im folgenden sollen nun noch einige allgemeine Anmerkungen zur Verwendung von Testbatterien gemacht werden:

In Teil I haben wir gesehen, daß die Veränderung von Intelligenztestleistungen über größere Zeiträume zwischen den Verbaltests geringer sind als solche zwischen anderen. Ferner haben wir aufgezeigt, daß ein Versuchsfehler möglicherweise ganz oder teilweise als Positionseffekt interpretiert werden kann. Auch bei unserer Kontrolluntersuchung scheint es notwendig zu sein, die Tatsache möglicher Positionseffekte als Folge einer immer wiederkehrenden Testsequenz etwas genauer zu betrachten: Für Testbatterien gilt allgemein, die Subtests immer in der im Manual vorgegebenen Reihenfolge durchzuführen, da in eben dieser Darbietungsfolge auch die Eichung der Testbatterien vorgenommen wurde, bei einer Veränderung der Reihenfolge also die Verwendung der Normwerte fragwürdig würde. Diese

Testfolge ist keine notwendige, sondern ergibt sich vermutlich eher aus der Tatsache, daß Testautoren kaum die personellen, finanziellen und technischen Mittel zur Verfügung stehen, um eine Testbatterie so zu eichen, daß mögliche Positionseffekte, die aus immer gleicher Testfolge resultieren, reduziert oder gar aufgehoben werden.

Die Abfolge der meisten in der BRD bekannten Testbatterien beginnt mit Verbaltests (z.B. HAWIE, IST 70 von AMTHAUER, P-S-B von HORN, ebenso die PMA 13-14), während nicht-verbale Tests erst danach folgen. Es könnte nun sein, daß Testergebnisse gerade bei Personen, die erstmals mit Tests konfrontiert werden, von deren Befangenheit, Angst oder Hemmung zu Beginn einer Untersuchung mitbestimmt werden, was für die Resultate in Verbaltests von besonderer Bedeutung wäre. Von daher ist denkbar, daß in dieser Hinsicht sensible Probanden systematisch benachteiligt werden, aber auch für verbal schwache Pbn könnte dies eine zusätzliche Behinderung bedeuten. Für die Testeichung sollte künftig ein Verfahren gefunden werden, das die oben angedeuteten möglichen Benachteiligungen beseitigt. Dies scheint uns auch im Hinblick auf die Diskussion über die zukünftige Entwicklung und Verwendung zusätzlicher Begabungstests bei der Vergabe von Studien- und Ausbildungsplätzen eine unabdingbare Notwendigkeit

zu sein. Außerdem erlaubt - streng genommen - nur eine auf Permutation der Einzeltests beruhende Eichung die Verwendung bestimmter Subtests und deren Normen, wenn also nicht die gesamte Batterie durchgeführt werden soll, wie es bei bestimmten Problemstellungen geschieht. So wird z.B. zur Diagnose von Legasthenikern zur Ermittlung der allgemeinen Intelligenz vielfach nur der nicht-verbale Teil einer Testserie verwendet (MACHEMER 1972, S. 90) wobei ja dann andere Tests diese Einstiegsfunktion übernehmen.

Bei Testwiederholungen finden sich häufig höhere Ergebnisse, die in aller Regel als Lernerfolg interpretiert werden. Es kann nun nicht ausgeschlossen werden, daß ein geringerer Lernerfolg bei Verbaltests gerade daher rührt, daß diese Tests am Anfang der Untersuchung stehen und sich die Pbn noch nicht in der gewünschten Entspanntheit befinden, die sich bei Bearbeitung der folgenden Tests allmählich einstellt. Ob und wie weit solche Effekte mittels "Eisbrechertests" reduziert oder vermieden werden können, ist u.W. einstweilen noch nicht systematisch geklärt. Wichtig scheint jedoch, darauf zu verweisen, daß sich Positionseffekte in der individuellen Diagnose nachteilig auswirken können. Für den Einsatz von bestimmten Formdeutungsverfahren (FDV) haben SPITZNAGEL und KORNADT verschiedene

Aspekte von Positionseffekten problematisiert. Wir teilen jedoch nicht die wohl zu optimistische Auffassung SPITZNAGELS: "Um soweit als möglich den Positionseffekt auszuschließen, muß der Testleiter die Beeinflußbarkeit des FDV durch vorausgehende Verfahren bedenken" (1964, S. 579). Dies kann ja wohl nur bedeuten, auf diese Weise sei das Problem zu lösen. Vielmehr schließen wir uns KORNAOT an, wenn er feststellt: "Die Frage, wie sich derartige Positionseffekte in der Individual-Diagnostik auswirken, ist noch völlig offen" (1964, S. 646) und generalisieren sie auch auf die Verwendung von Intelligenztests. Die Ergebnisse unserer Kontrolluntersuchung werden auch unter diesem Aspekt diskutiert werden müssen. Dabei ist außerdem zu berücksichtigen, daß die Eichstichprobe für das P-S-B von HORN nur für Baden-Württemberg repräsentativ ist (AURIN 1968, S. 34).

Auf die Versuchsvorbereitung, Durchführung, Auswertung, etc. soll hier nicht näher eingegangen werden, nur soviel: Soweit irgend möglich, haben wir uns in allen Phasen der Untersuchung an die Verhältnisse der Längsschnittuntersuchung angepaßt, um zusätzliche Varianz zu vermeiden. Es wurden immer zuerst die PMA, dann das PSB und zum Schluß der FAT dargeboten.

12.0 ERGEBNISSE DER KONTROLLUNTERSUCHUNG

Da an anderer Stelle über die mathematisch-statistische Analyse der Befunde einige allgemeine Ausführungen über die verwendeten Verfahren gemacht werden sollen, beschränken wir uns hier auf die Darstellung der Ergebnisse.

12.1 Prüfung auf Normalverteilung

Jeder Einzeltests wurde auf Normalverteilung geprüft. Dabei wurden die Rohwertpunkte zu Klassen zusammengefaßt und jeweils folgende Prüfgrößen ermittelt:

- a) χ^2 für die "Güte der Anpassung"
- b) t-Wert für die Schiefe der Verteilung
- c) t-Wert für den Exzeß der Verteilung

Die Normalverteilungshypothese wurde dann verworfen, wenn mindestens eine der drei Prüfgrößen auf dem 1%-Niveau signifikant war (LIENERT 1967, S. 171ff).

Wie aus Tabelle 56 ersichtlich, mußten aus beiden Testbatterien je zwei Tests als nicht normalverteilt gewertet werden. Auffällig ist, daß es jeweils analoge Tests sind, d.h. solche, die in beiden Batterien gleiche resp. ähnliche Merkmale messen

sollen, nämlich Figurenvergleich und Raumvorstellung sowie Rechentest und Addieren! Da die Prüfgrößen Aufschluß über die Art der Abweichung geben, seien sie in Tabelle 57 aufgeführt.

Tabelle 56 : Zusammenstellung der Normalverteilungsprüfungen (NV)

<u>PMA 13-14</u>	<u>NV</u>
Wortschatz	+
Figurenvergleich	-
Wahrnehmungsschnelligkeit	+
Buchstabenreihen	+
Wortgruppen	+
Rechentest	-
<u>PSB-HORN</u>	
Allgemeinbildung	+
Denkfähigkeit (3)	+
Denkfähigkeit (4)	+
Worteinfall	+
Ratefähigkeit	+
Raumvorstellung	-
Gliederungsfähigkeit	+
Addieren	-
Wahrnehmungstempo	+
FAT 7-8	+

"+" = Beibehaltung der Normalverteilungshypothese

"-" = Verwerfen der Normalverteilungshypothese

Tabelle 57 : Prüfgrößen von der Normalverteilung abweichender Tests

Test	chi ²	Schiefe t-Wert	Exzeß t-Wert
Figurenvergleich (PMA)	20,09 ++	-0,50 ns	-2,43 +
Raumvorstellung (HORN)	47,15 +++	-2,89 ++	-1,19 ns
Rechentest (PMA)	17,26 ++	4,15 +++	4,80 +++
Addieren (HORN)	10,92 ns	3,51 +++	3,51 +++

+++ = signifikant auf dem 0,1%-Niveau

++ = signifikant auf dem 1%-Niveau

+ = signifikant auf dem 5%-Niveau

ns = nicht signifikant

Die χ^2 -Tests ergaben sehr bedeutsame Unregelmäßigkeiten für Figurenvergleich, Raumvorstellung und Rechentest. Der Test Raumvorstellung ist zudem noch sehr signifikant rechtsschief, die Tests Rechentest und Addieren sind dagegen linkschief. Außerdem sind die zuletzt genannten Tests sehr signifikant schmalgipflig.

12.2 Validität der Tests

12.2.1 Interkorrelationen der PMA 13-14

Die Prüfung auf Normalverteilung hat nicht für alle Subtests die Normalverteilungshypothese bestätigt. Daher wenden wir zur Berechnung der Interkorrelationen

einheitlich den Rangtest (mit Korrektur bei Rangplatzverteilung) nach SPEARMAN an (SIEGEL 1956). Gleichzeitig führen wir zu Vergleichen auch die Interkorrelationen aus der Längsschnittuntersuchung (N = 212, alle Haupt-, Real- und Oberschüler) auf, Tabelle 58 .

Tabelle 58 : Rangkorrelationen der Einzeltestroh-
werte sowie der Gruppenstandardwerte
der PMA 13-14 für die Querschnitt-
stichprobe (N = 116) und die Längs-
schnittstichprobe (N = 212, ohne Son-
derschüler)

<u>Längsschnitt. N = 212</u>	<u>Kontrollstichprobe/N = 116</u>								
	1	2	3	4	5	6	7	8	9
1. Wortschatz		27	16	33	45	18	49	33	58
2. Figurenvergleich	11		22	31	33	03	36	70	61
3. Wahrnehmungsschnelligkeit	23	32		53	21	01	43	64	55
4. Buchstabenreihen	58	34	46		40	21	82	56	71
5. Wortgruppen	64	23	25	60		08	83	35	65
5. Rechentest	28	28	35	38	35		19	47	42
7. Reasoning 13-14	68	31	39	88	89	39		53	82
8. Lesefrei 13-14	28	72	73	52	37	71	50		85
9. GSW 13-14	68	54	61	81	75	62	87	82	

Bestimmt man über die z' -Transformation für die 6 Einzeltests jeder der beiden Korrelationsmatrizen die mittlere Interkorrelation, so wird deutlich, daß der mittlere Zusammenhang zwischen den Tests der

Die Koeffizienten wurden mit 100 multipliziert.
Für N=116 entsprechen $r > 17$ einem $p < 5\%$; $r > 22$ einem $p < 1\%$.
Für N=212 entsprechen $r > 13$ einem $p < 5\%$; $r > 18$ einem $p < 1\%$.

Längsschnittstichprobe mit $r_m = 0,37$ größer ist als zwischen denen der Querschnittstichprobe mit $r_m = 0,25$. Die Differenz dieser Koeffizienten ist hoch signifikant, d.h., bezogen auf die zugrunde liegenden Testvariablen sind die beiden Probandengruppen keine Zufallsstichproben einer gemeinsamen Grundpopulation.

Die Annahme der Zufallsauswahl war notwendig für die Querschnittstichprobe, sofern weiterreichende Schlüsse aus den Ergebnissen gezogen werden sollen. Natürlich gibt die Signifikanz der Differenz der Koeffizienten keinen Aufschluß darüber, welche der beiden Untersuchungsgruppen von der empirischen Gesamtpopulation aller Haupt-, Real- und Oberschüler der Bundesrepublik abweicht. bzw. mit ihr übereinstimmt.

Eine Prüfung analoger Interkorrelationen der sechs Einzeltests beider Stichproben (z.B. Wortschatz mit Figurenvergleich, $r_{116} = 0,27$, $r_{212} = 0,11$) auf Überzufälligen Unterschied ergab mit Ausnahme der Interkorrelationen von Buchstabenreihen mit Figurenvergleich, sowie Wahrnehmungsschnelligkeit mit Wortgruppen nach den oben dargestellten Ergebnissen erwartungsgemäß mindestens bedeutsame Unterschiede. Auch dies ist ein Indiz zur Mahnung vor vorschnellen Generalisierungen.

12.2.2 Interkorrelationen der PSB

Tabelle 59 zeigt die Interkorrelationen des PSB und zwar diejenigen des Manuals und die unserer Stichprobe. Vergleicht man beide Matrizen miteinander, so kann man befriedigende Übereinstimmungen konstatieren, die am ähnlichsten in den Tests "Denkfähigkeit" ist. Auch die Übereinstimmung der Korrelationen mit Gl ist mit Annahme der "Allgemeinbildung" befriedigend. Man kann also sagen, daß die beiden Stichproben in ihren Strukturen einander ähnlich sind.

Tabelle 59 : Interkorrelationen (SPEARMAN) des PSB von HORN
Über der Diagonalen: Aus dem PSB-Manual N = 907
unter der Diagonalen: Kontrollstichprobe N = 116

	1+2	3	4	3+4	5	6	7	8	9	10	GL
1+2		.41	.45	.47	.53	.68	.31	.39	.43	.36	.75
3	.24		.63	.89	.35	.39	.36	.42	.33	.34	.65
4	.35	.60		.88	.37	.42	.37	.40	.35	.34	.66
3+4	.32	.89	.88		.38	.44	.39	.45	.37	.37	.72
5	.37	.28	.30	.31		.43	.25	.34	.40	.33	.63
6	.60	.31	.27	.32	.36		.26	.36	.37	.31	.69
7	.17	.47	.36	.47	.19	.26		.37	.25	.34	.60
8	.30	.42	.31	.41	.28	.40	.41		.29	.30	.63
9	.14	.41	.32	.41	.29	.15	.24	.27		.37	.59
10	.06	.31	.29	.32	.16	.21	.23	.24	.42		.64
GL	.55	.69	.62	.73	.54	.57	.63	.63	.57	.55	

1+2 = Allgemeinbildung 3, 4, 3+4 = Denkfähigkeit
5 = Worteinfall 6 = Ratefähigkeit
7 = Raumvorstellung 8 = Gliederungsfähigk.
9 = Addieren 10 = Wahrnehmungstempo,
GL = Gesamt1.

12.3

Faktorenanalyse Über 6 PMA 13-14,
9 PSB-Tests und FAT 7-8

Zur weiteren Absicherung der Übereinstimmungsvalidität der beiden Test-Batterien PMA und PSB und Über den FAT 7-8 haben wir Über die Standardwerte der insgesamt 16 Tests eine Faktorenanalyse gerechnet; Stichprobengröße $N = 116$. Faktorenanalysen Über homogenere Stichproben zu rechnen, waren wegen der verbleibenden Gruppengrößen nicht sinnvoll. Tabelle 61 enthält die Ergebnisse.

Am ehesten war die 6er Lösung zu interpretieren. Dabei klärt der höchste Faktor 25% und der niedrigste immer noch 8,2% der gemeinsamen Varianz auf. Hauptinteresse an der Faktorenanalyse war die Frage, ob die Tests der Testbatterien, die jeweils für bestimmte Faktoren sensu THURSTONE stehen, auf denselben Faktoren laden, so daß möglichst die THURSTONEschen Faktorenbezeichnungen verwendet werden könnten. Die Ladungen können als mittelhoch bezeichnet werden. Am deutlichsten können die Faktoren II und IV interpretiert werden. Im ersten Fall handelt es sich um den Faktor "Number", im zweiten um "Perceptual Speed". Faktor V ist der "Reasoning-Factor". Daß der FAT hier läßt, entspricht der Konzeption dieses Tests, allerdings hätten die Autoren BELSER/ANGER/BARGMANN (1965)

Tabelle 61 : Faktorenanalyse Über 6 PMA 13-14
und 10 PSB-Tests

N = 116							
	Faktoren						Kommuni- täten
Testvariablen	I	II	III	IV	V	VI	
FAT 7-8	.59		.45		.44		.76
Allgemeinbildung	.85					.28	.85
Denkfähigkeit I			.47		.55		.65
Denkfähigkeit II			.25		.70		.68
Wortefinfall					.25	.54	.44
Ratetfähigkeit	.59		.26	.33		.45	.72
Raumvorstellung			.61		.20		.47
Gliederungsfähigkeit			.52				.41
Addieren		.87					.85
Wahrnehmungstempo		.29		.75			.69
Wortschatz	.77						.66
Figurenvergleich			.67				.52
Wahrnehmungsschn.				.58	.41	.35	.70
Buchstabenreihen					.70		.69
Wortgruppen	.55				.39		.53
Zahlen		.78					.64
Varianzanteile (%)	25.0	16.5	18.5	12.0	19.8	8.2	=100

Ladungen unter .25 wurden der Übersichtlichkeit wegen fortgelassen.
Die Faktorenanalyse wurde mit dem Programm PAFA von SCHNELL & GEBHARDT Darmstadt in der Universität Münster - wie alle anderen Berechnungen auch - gerechnet. Den Mitarbeitern des Rechenzentrums der Universität sei für ihre Hilfe gedankt.

sicher eine deutlichere Ausprägung erwartet. Neben den Reasoning-Tests läßt noch der PMA-Test Wahrnehmungsschnelligkeit auf diesen Faktor. Als relativ deutlich kann auch der Faktor I, "Verbal", interpretiert werden. Er vereinigt auf sich die Verbal-Tests und die beiden Reasoning-Tests, die eine klare verbale Komponente haben. Der Faktor III entspricht dem "Space-Factor". Die Ladungen in FAT und Denkfähigkeit I ergänzen das Bild, handelt es sich doch um Ladungen, die hauptsächlich auf die räumlich-geometrischen Figuren dieser beiden Tests zurückgehen. Den Faktor VI zu benennen, macht einige Schwierigkeiten. Es sind zwei Verbal-Tests, wobei der Wörter-Ratetest interessanterweise keine substantielle Ladung auf dem Faktor I hat. Wir wollen den Faktor VI "Verbal Fluency" (sensu THURSTONE) nennen, weil wir meinen, daß die Tests fordern, sich unter Zeitdruck viele Wörter einfallen zu lassen bzw. "zerbrochene" und mit Denkfehlern versehene Wörter richtig zu erkennen, Fähigkeiten also, die Bestandteil von Wortgewandtheit sein dürften.

Wir können also zusammenfassend sagen, daß eine Übereinstimmung der beiden Tests gegeben ist. Sie ist aber mit einigen Mängeln behaftet.

12.4

Mittelwertunterschiede intra Kontrolluntersuchung und inter Kontroll- und Längsschnittuntersuchung

Als wir die Kontrollstichprobe konzipierten, hatten wir die Erwartung, daß etwa die Längsschnittstichprobe und die Kontrollstichprobe, ebenso aber auch die nach demographischen Merkmalen gleichen Substichproben sich jeweils nicht unterscheiden würden, zumindest sollten systematische Effekte zwischen Längsschnitt- und Kontrollstichprobe nicht vorkommen. Solche systematischen Effekte würden entweder auf einen Bias, der entweder stichprobenbedingt oder situationsbedingt sein müßte, hinweisen, würde also vor Generalisierungen warnen.

a) Mittelwertunterschiede der Kontrollstichprobe

In der Tabelle 62 geben wir eine Übersicht über mindestens signifikante Stichprobenunterschiede. Wie zu erwarten, sind in allen signifikanten Mittelwertunterschieden die OS besser als die RS und die RS wiederum erzielen bessere Ergebnisse als die HS.

Bei den Geschlechtern sind mit einer Ausnahme - Figurenvergleich - immer die Mädchen (MA) besser als die Jungen (JU). Zum großen Teil beziehen sich die signifikanten Unterschiede jeweils auf analoge Tests der Testbatterien.

Tabelle 62 : Vergleiche zwischen HS,RS,OS und Jungen und Mädchen mit dem U-Test von MANN & WHITNEY

Testvariable	HS/RS	HS/OS	RS/OS	JU/MA
FAT 7-8	5	0,1		
Allgemeinbildung	5	1		
Denkfähigkeit I		1		
Denkfähigkeit II	5	0,1	1	
Reasoning	5	0,1		
Wortefall				1
Ratfähigkeit				
Raumvorstellung				
Gliederungsfähigkeit	5	1		
Addieren				5
Wahrnehmungstempo			5	1
Gesamtleistung +	5	5		5
Wortschatz		0,1	0,1	
Figurenvergleich				1
Wahrnehmungsschn.			1	0,1
Buchstabenreihen	5	0,1	1	1
Wortgruppen	5	1		
Zahlen	5			
Reasoning +	1	0,1	1	
Lesefrei +				
GSW 13-14 +	5	0,1	5	5

Die Signifikanzangabe steht immer unter der Gruppe, die die HÖHERE Testleistung aufweist.

Bei den Variablen mit + liegen der Berechnung Standardwerte zugrunde, bei den übrigen Rohwerte.

b) Mittelwertunterschiede zwischen analogen Gruppen
der Kontroll- und der Längsschnittstichprobe

In diesem Teil der Analyse haben wir jeweils die
Testleistungen in den PMA 13-14 der HS, RS, OS der
Kontrolluntersuchung mit den analogen Gruppen der
Längsschnittuntersuchung verglichen. Die Ergebnisse
waren ein wenig Überraschend. Keinerlei Unterschiede
gab es bei den Tests

Figurenvergleich

Buchstabenreihen

Wortgruppen,

Reasoning.

Keine signifikanten Unterschiede finden sich außer-
dem im Test

Wortschatz bei HS und OS,

Wahrnehmungsschnelligkeit bei HS,

Zahlen bei RS,

Lesefrei bei HS,

GSW 1968 bei HS.

Alle verbleibenden Unterschiede sind signifikant.

In allen Fällen erzielten die Schüler der Längs-
schnittstichprobe bessere Leistungen als diejeni-
gen der Kontrollstichprobe; es sind dies im Test

Wortschatz die RS,

Wahrnehmungsschnelligkeit die RS, OS,

Zahlen die HS, OS,
Lesefrei die RS, OS,
GSW 1968 die RS, OS.

Worin könnten die systematischen Signifikanzen ihre Ursachen haben. Eine Entscheidung scheint nicht möglich. Aber wenigstens drei mögliche Gründe können angeführt werden:

Der naheliegendste Grund könnte in von uns nicht kontrollierten Merkmalen der Kontrollstichprobe zu suchen sein. Daß diese Möglichkeit überhaupt ernsthaft in Betracht zu ziehen ist, verweist einmal mehr darauf, alle untersuchungsrelevant erscheinenden Merkmale von Stichproben - zumindest, wenn es sich nicht nach Zufall auf bestimmte experimentelle Untersuchungen aufgeteilte Probanden handelt - zu ermitteln und mitzuteilen. Der zweite Grund könnte in den Durchführungsbedingungen liegen. Es wäre denkbar, daß Versuchsleiter und -helfer (Studenten der Psychologie) trotz praktischer Einarbeitung nicht hinreichend trainiert waren. Der dritte Grund ist die zeitliche Differenz der beiden Untersuchungen. Die Kontrolluntersuchung wurde 1970, also zwei Jahre nach der Längsschnittstudie durchgeführt. Damit gehören die Schüler der Kontrolluntersuchung zu einer Population, die stärker von der Bildungswerbung angesprochen worden sein mag. Dies könnte dann auch mehr schwächere Schüler zur RS oder OS geführt haben.

12.5

Regressionsanalysen über GL-PSB und GSW 13-14

In einem weiteren Abschnitt wollen wir mit Hilfe der Regressionsanalyse untersuchen, wie genau die Gesamtergebnisse eines Tests aus dem jeweils anderen geschätzt werden können. In der Tabelle 63 sind die wichtigsten Daten zusammengefaßt.

Tabelle 63 : Daten zu den Regressionsanalysen

Variablen	AM	ST	Regr.koeff.	r
GL-PSB	50,6	8,2	0,70	0,77
GSW 13-14	48,5	7,5	0,84	

Regressionsgleichungen:

$$(a) Y_{GL-PSB} = 0,70 X_{GSW\ 13-14} + 13,13$$

$$(b) Y_{GSW\ 13-14} = 0,84 X_{GL-PSB} + 9,84$$

Standardschätzfehler für

$$(a) s_e = 4,83$$

$$(b) s_e = 5,30$$

Die Höhe der Regressionskoeffizienten ist mit 0,70 und 0,84 zufriedenstellend. Mit Hilfe dieser Gleichungen ist nun möglich abzuschätzen, welchen GSW ein Proband im PSB haben würde, wenn sein GSW im

PMA 13-14 bekannt ist und umgekehrt. Wie man sich leicht überzeugen kann, sind die Differenzen im oberen GSW-Bereich am größten, so ist bei einer Schätzung, die wechselweise von einem Standardwert = 70 ausgeht, schon eine Differenz von 6,5 Standardwertpunkten anzusetzen.

12.6 Veränderungsmessung kognitiver Merkmale bei wiederholter Messung

Im vorigen Abschnitt hatten wir versucht, Gründe für systematische Unterschiede zwischen der Längsschnittstichprobe und der Kontrollstichprobe zu finden. Wir wollen dieses Problem hier unter einem etwas anderen Aspekt nochmals aufnehmen und fragen, ob sich in der Kontrollstichprobe unterschiedliche Leistungen zeigen, wenn wir "analoge" Tests der verschiedenen Testbatterien miteinander vergleichen.

a) Veränderungen der Kontrollstichprobe

In der Tabelle 64 haben wir die Mittelwertabweichungen zusammengetragen, die sich zwischen "analogen" Tests der Gesamtstichprobe ergeben haben. Es zeigen sich sehr signifikante Veränderungen zwischen den Reasoning-Tests, aber auch zwischen den Gesamtleistungstests. Auffällig ist vor allem, daß die Veränderungen die Reasoning-Tests am meisten betreffen.

Tabelle 64 : Differenzen "analoger" T-Standard-
werte (WILCOXON-Test für korre-
lierte Stichproben)
Normmittelwert = 50

Testdifferenzen zwischen ...	alle Schüler N = 116		p%
	AM	AM	
FAT/PSB-Reasoning	52,8	53,6	
FAT/PMA-Reasoning	52,8	50,1	1
PSB-Reas./PMA-Reas.	53,6	50,1	1
PSB-Allg./Wortsch.	50,3	49,9	
PSB-Add./Zahlen	45,8	47,2	
PSB-Gl./GSW 13-14	50,6	48,5	1
FAT/Wortschatz	52,8	49,9	1
FAT/GSW 13-14	52,8	48,5	1
PSB-Reas./PSB-Allg.	53,6	50,3	1
PSB-Allg./GSW 13-14	50,3	48,5	1
PSB-Reas./PSB-Gl	53,6	50,6	1
PSB-Reas./GSW 13-14	53,6	48,5	1

p% entspricht der 1% Irrtumswahrscheinlichkeit
AM = Arithmetischer Mittelwert

Darüberhinaus sind die "Zugewinne" immer auf Seiten der PSB und FAT. Wie kann man dieses deuten? Wie bekannt, wurden die Tests in der Reihenfolge PMA, PSB und FAT durchgeführt. Einmal können die Ergebnisse alle mit der Regression zur Mitte verstanden werden, wenn man die Mittelwerte vergleicht. Es macht aber stutzig, daß hier wiederum die Reasoning-Skores so deutlich beteiligt sind. Wir stellen hier die Frage, ob dies damit zu tun haben könnte, daß es bei Reasoning-Tests mehr als bei anderen darum

geht, das generelle Lösungsschema zu finden, was sich im Erfolgsfalle in "temporär" höheren Leistungen auszahlt. Oder sind dies einfach die Tests, die ad hoc Lernen erfordern, wobei derjenige besonders gut abschneidet, für den Lernen quasi Lebensprinzip ist? Es wäre empfehlenswert, wenn auf diesen Testtyp, die ja alle "Denken" messen sollen, die Untersuchungsmethoden der Denkpsychologie verstärkt Anwendung fänden (SIMONS & LANGHEINRICH 1980).

Ein weiterer Grund dafür, daß die Leistungen in den PMA immer unter den analogen der anderen Tests lagen, kann zwei weitere Gründe haben: Das PSB könnte leichter sein als die PMA (vgl. auch KEMMLER 1976, S. 60). Andererseits könnte hier ein positiver Lerntransfer stattgefunden haben, da die PMA immer als erste Tests durchgeführt wurden. Streng genommen, sicher ein Versuchsfehler. Wir würden auch - nicht zur eigenen Entlastung - die üblichen immer gleichen Reihenfolgen bei Testeichungen für einen Methodenfehler halten, der zumindest systematische Bias zu produzieren geeignet ist (in Bezug auf Normen).

b) Intra-individuelle Differenzen

Wie in der Längsschnittuntersuchung wollen wir auch hier die Frage nach individuellen Differenzen, die größer sind als das Konfidenzintervall zweier Messungen, stellen.

Tabelle 65 enthält eine Aufstellung mit 12 Probanden, deren IQ sich um mindestens eine Standardeinheit verändert haben. Gebildet wurde die Differenz zwischen der standardisierten Gesamtleistung des PSB minus dem GSW 13-14. Insgesamt schwanken die Differenzen zwischen 15-22 IQ-Punkten. Von den Probanden sind vier Real- und sieben Hauptschulmädchen. Alle Differenzen sind positiv. Die einzige negative Differenz von -19 IQ-Punkten stammt von einem Hauptschuljungen. Sowohl das Verhältnis der Geschlechter als auch das Vorzeichen der Differenz weichen sehr stark von der Längsschnittuntersuchung ab. Erklärungsgründe dafür sind kaum anzugeben. Es fällt aber auf, daß die Ausgangslage im GSW 13-14 bei 10 von 12 Probanden eher niedrig war, so daß mit einem Regressionseffekt zur Mitte gerechnet werden mußte. Es kann sich daher sicher auch kaum um echte Gewinne oder Verluste handeln, so daß man diese Schüler auch nicht mit Auf- oder Absteiger zutreffend kennzeichnen würde.

Tabelle 65 : Intra-individuelle Differenzen
zwischen G1-HORN - GSW 13-14
(in IQ-Werten)

Pb-Nr.	G1-HORN	GSW 13-14	IQ-Diff.	Schüler
16	86	104	-19	HSm
13	115	100	+15	HSw
26	95	76	+19	HSw
9	95	77	+18	HSw
20	101	79	+22	HSw
30	100	85	+15	HSw
32	115	100	+15	HSw
45	112	92	+20	HSw
119	97	82	+15	RSw
122	125	104	+21	RSw
117	112	92	+20	RSw
127	115	95	+20	RSw

Die Transformation in IQ-Differenzen soll die doch erheblichen Veränderungen deutlich machen.

12.7 Diskriminanzanalysen

Wir wollten die D-Analyse auch bei dieser Stichprobe anwenden, um zu sehen, welche Testbatterie der faktisch von den Schülern getroffenen Schullaufbahn am nächsten komme, oder ob möglicherweise beide Testbatterien zu ganz anderen, völlig abweichenden Klassifikationen kommen würden. Den FAT haben wir hierbei ausgeschlossen.

a) D-Analyse der HS, RS, OS, über 6 PMA 13-14

Tabelle 66 zeigt die Zuordnungen; das D^2 ist sehr signifikant.

Tabelle 66 : Zuordnung nach Diskriminanzanalyse
6 PMA 13-14, Stichproben HS, RS, OS
 $d^2 = 82,9$ $df = 12$

Zuordnung nach D-Analyse				
	HD	RD	OD	Su.
HS	43	11	4	58
RS	6	19	5	30
OS	2	6	20	28
Su.	51	36	29	116

Danach werden 43 HS (= 74%), 19 RS (= 63%) und 20 OS (= 71%) zusammen also 83 Schüler (= 71%) entsprechend ihrer tatsächlichen Schulart zugeordnet. 14 Schüler (= 12%) werden als Überforderte und 20 (= 17%) als Unterforderte oder sog. "Begabungsreserve" klassifiziert.

b) D-Analyse der HS, RS, OS, über 9 PSB

Tabelle 67 zeigt die Zuordnungen; das D^2 ist sehr signifikant.

Tabelle 67 : Zuordnung nach Diskriminanzanalysen
9 PSB, Stichproben HS, RS, OS

$$D^2 = 60,2 \quad df = 18$$

Zuordnung nach D-Analyse				
	HD	RD	OD	Su.
HS	35	15	8	58
RS	8	18	4	30
OS	3	6	19	28
Su.	46	39	31	116

Nach der Klassifikation sind 35 HS (= 60%), 18 RS (= 60%) und 19 OS (= 68%) zusammen 72 Schüler (= 62%) schulbesuchsgemäß zugeordnet, also 9% weniger als bei den PMA. Entsprechend verändern sich auch die Zahlen der Überforderten, 17 Schüler (= 15%) und der Unterforderten, 27 Schüler (= 23%). Hier ist also sogar eine um 5% höhere sog. "Begabungsreserve" vorhanden.

Wenn man auch hier (s. oben Teil I) auszählt, wie Übereinstimmend in beiden D-Analysen die Klassifikation erfolgt ist, so ergibt sich folgendes Bild:

64 Schüler (= 55%) werden bei beiden D-Analysen entsprechend ihrer tatsächlichen Schulart zugeordnet; bei 39 Schülern (= 34%) erfolgt die Abweichung um jeweils eine Schulart. Und 13 Schüler (= 11%) werden in der einen D-Analyse z.B. zur HS und in der anderen zur OS klassifiziert und vice versa.

Für die Praxis wäre es jedoch einfacher, beide Testbatterien in eine D-Analyse einzubeziehen. Wir haben dies ebenfalls getan. Tabelle 68 zeigt die Klassifikationen.

Tabelle 68 : Zuordnung der Diskriminanzanalyse
6 PMA 13-14 und 9 PSB,
Stichprobe HS, RS, OS

$$D^2 = 140,0 \quad df = 30$$

Zuordnung nach D-Analyse				
	HD	RD	OD	Su.
HS	46	9	3	58
RS	6	20	4	30
OS	3	1	24	28
Su.	55	30	31	116

Wie zu erwarten, werden nunmehr 46 HS + 20 RS + 24 OS = 90 Schüler (= 78%) ihrer Schulart spezifisch zugeordnet. Immerhin zählen wir 6 RS + 4 OS = 10 Schüler (= 9%), die nach unserer Definition überfordert und 12 HS + 4 RS = 16 Schüler (= 14%) unterfordert sind, also als sog. "Begabungsreserve" zur Verfügung stünden.

Die Zahlen dürfen nicht absolut gesetzt werden. Sie beziehen sich auf die Annahme der kognitiven Struktur als dem wesentlichsten Kriterium für Schulerfolg.

Und sie beziehen sich auf unsere Stichprobe und die verwendeten Tests. Außerdem beziehen sich die Zahlen auf das Jahr 1970. Daß dies jedoch kein Nachteil für uns zu sein braucht, kann mit dem Hinweis angedeutet werden, daß die Stadt Münster - wie bereits erwähnt - schon am Ende der 60er Jahre Schülerquoten zur QS und RS schickte, die sehr weit über dem Durchschnitt NRW's und des Bundesgebietes lagen. Einige Zahlen sollen dies noch einmal belegen:

Von 315 Schülern gingen 79 ursprünglich zur QS (= 25%), 50 zur RS (= 16%). Dies sind Mindestzahlen, da Schulwechsler nicht immer erfaßt werden konnten. 37 (= 12%) von 315 oder 47% von 79 machten Abitur, auch dies nur Mindestzahlen (vgl. FIEGENBAUM 1974, S. 216; Bildungsbericht 1970, S. 51).

13.0

ZUSAMMENFASSENDE INTERPRETATION DER
ERGEBNISSE DER KONTROLLUNTERSUCHUNG

Die Kontrolluntersuchung ist als Kreuzvalidierung zur Längsschnittuntersuchung angelegt. Es galt vor allem, die Frage zu prüfen, ob die im Rahmen der Längsschnittuntersuchung ins Deutsche Übersetzte und zusammengestellt Testbatterie PMA 13-14 ein valides Instrument sei.

Dazu wurden die PMA 13-14 zusammen mit den geeigneten Tests PSB von HORN und FAT 7-8 von BELSER et al. erhoben. Das PSB wurde gewählt, weil HORN mitteilt, es nach der Faktorentheorie THURSTONE's konstruiert zu haben. Ebenso erklären BELSER et al. in der Handanweisung, ihr FAT messe THURSTONE's Faktor Reasoning, was also eine zusätzliche Kontrollmöglichkeit eröffnete, zumal wir dem Faktor Reasoning (neben Verbal) besondere Bedeutung für die Schullaufbahnentwicklung beimessen.

Die Stichprobe sollte in ihrer Zusammensetzung der Längsschnittstichprobe gleichen. So ergab sich schließlich ein $N=116$ Schüler = 58 HS + 30 RS + 28 OS. Das Durchschnittsalter betrug 14 Jahre, die Untersuchung wurde an Münsteraner Schulen im Rahmen der normalen Schulstunden durchgeführt.

Durchführung und Auswertung der Untersuchung wurden nach Einarbeitung durch den Autor von Studenten der Psychologie geleistet.

Ergebnisse:

Sowohl von den PMA 13-14 als auch den PSB waren die Tests die die Faktoren "Number" und "Space" messen sollen, nicht normalverteilt, allerdings nicht alle gleichsinnig.

Zur Validitätsprüfung haben wir dann verschiedene Rankkorrelationen berechnet bzw. Vergleiche durchgeführt:

- Interkorrelationen der PMA über die Kontrollstichprobe,
- Interkorrelationen der PMA über die Längsschnittstichprobe,
- Interkorrelationen des PSB über die Kontrollstichprobe,
- Interkorrelationen des PSB über die Eichstichprobe (Manual).

Die Korrelationsmatrizen sind zwar einander ähnlich, man hätte im Sinne eines Paralleltests jedoch noch eine stärkere Ähnlichkeit erwartet. Schließlich haben wir auch noch die Testergebnisse aller PMA, PSB und FAT untereinander korreliert. Auch hier sind die erwarteten Koeffizienten nur mittelhoch.

So korrelieren der GSW 13-14 mit G1-HORN nur zu 0,74 (vgl. MANDL 1975, S. 73) und mit dem FAT nur 0,69. G1-Horn und FAT korrelieren 0,68. Zur besseren Übersicht der zuletzt genannten Korrelationsmatrize haben wir eine Faktorenanalyse über N=116 und alle 6 PMA, 9 PSB und den FAT gerechnet. Ohne Ausnahme konnten wir auf diese Weise die THURSTONE-sche Faktorenstruktur soweit sie mit diesen Tests erfaßt wird, replizieren.

Als nächstes haben wir Mittelwertvergleiche durchgeführt. Bei den Vergleichen unkorrelierter Stichproben sind die Vergleich immer zugunsten der "höheren" Schule ausgefallen. Beim Vergleich der Geschlechter haben die Mädchen in sieben von acht signifikanten Werten besser abgeschnitten als die Jungen.

Wir haben dann auch die jeweils analogen Schulartgruppen der Längsschnitt- und der Kontrollstichprobe miteinander verglichen. Von den 27 Vergleichen zeigten 18 keine signifikanten Unterschiede. Die 9 signifikanten Unterschiede waren alle zugunsten der Längsschnittuntersuchung. Ursachen für die Differenzen und deren Einseitigkeit können sowohl in der Stichprobenzusammensetzung (!) als auch in den situativen Bedingungen zu suchen sein.

Als letzte Prüfung der Übereinstimmungsvalidität haben wir auch noch die beiden Regressionsgleichungen für die Gesamttestwerte berechnet, wobei beide Werte wechselseitig unabhängige bzw. abhängige Variable sind. Auch diese Prüfung ist mit Regressionskoeffizienten von 0,70 und 0,84 zufriedenstellen ausgefallen.

Obgleich es nicht zu den üblichen Prozeduren einer Validitätstudie gehört, haben wir noch den Effekt wiederholter Messung untersucht. Wir stellten fest, daß die Wiederholungseffekte immer zugunsten unserer Referenztests ausfielen. Hierbei kann es sich um einen Lerneffekt handeln, oder die PMA sind schwerer als die anderen Tests.

Abschließend haben wir auch hier nochmals untersucht, welche Möglichkeiten die Diskriminanzanalyse bei der Schullaufbahnberatung spielen kann. Wir erhielten bei beiden Testbatterien klare Ergebnisse, die fragen lassen, warum die Diskriminanzanalyse (u.W.) noch immer die bekannte Abstinenz erfährt.

Alle Ergebnisse zusammenfassend, kann gesagt werden, daß die Kontrolluntersuchung Informationen erbracht hat, die hinter dem, was in der Literatur üblich ist, nicht zurückzustehen brauchen. Das Verfahren der

Übereinstimmungsvalidität hat allerdings methodologische Grenzen. Es basiert auf einem prinzipiell infiniten Regreß, und führt sich damit sehr leicht selbst ad absurdum, nämlich dann und insofern die Validität des jeweiligen Vergleichstests angenommen wird oder werden muß, ohne tatsächlich nachgewiesen zu sein.

14.0

ZUSAMMENFASSENDE INTERPRETATION
DER ERGEBNISSE DIESER ARBEIT
UND AUSBLICK

Zunächst wollen wir unsere Fragestellungen in dem Kontext, in dem sie sich uns stellten, wiederholen: Die vorliegende Arbeit besteht aus zwei Teilen. Der erste Teil basiert auf Daten einer 1960 von KEMMLER begonnenen und 1974 abgeschlossenen Längsschnittstudie zur Bedingungsanalyse von Schulerfolg und Schulversagen (KEMMLER 1967, 1976). Es handelt sich bei diesem Unternehmen um einen "großangelegten Erkundungsversuch" (KEMMLER 1967, S. 13). Dieser Sachverhalt beeinflusst auch unsere Fragestellungen. Sie sind - entsprechend der Konzeption der Längsschnittstudie - explorativ formuliert. Gezielte Hypothesenformulierungen und -prüfungen können aus wissenschaftsethischen Gründen hier nicht erfolgen, sie hätten u. E. eine entsprechende Anlage der Untersuchung erfordert.

Während der zeitweiligen Mitarbeit an der Längsschnittuntersuchung wurden wir u. a. mit methodologischen Problemen konfrontiert, die in der Forschungspraxis (Publikationen) nicht selten geringgeachtet werden, obgleich sie für die Ergebnisse einer Arbeit von besonderer Bedeutung sind. Wir haben die seltene Chance, Daten einer Längsschnittuntersuchung verwenden zu können, gerne wahrgenommen, um einige der Fragen zu bearbeiten.

Unsere erste Frage war, welche Informationen über die Stichprobe einer Untersuchung notwendig seien, um der methodologischen Forderung nach Replizierbarkeit einer Untersuchung und der Vergleichbarkeit von Befunden zu entsprechen.

Das Problem mangelhafter Stichprobenauswahl brachte uns auf die Frage nach dem Einfluß, den dies auf die instrumentellen Eigenschaften von Tests haben müsse. Wir haben uns die Frage gestellt, ob denn die Eichstichproben einiger der bekanntesten deutschsprachigen Intelligenztests dem Kriterium der Repräsentativität der Eichstichproben genügen, und welche Konsequenzen gegebenenfalls aus unseren Befunden abzuleiten seien.

Eine weitere Frage galt der Veränderung von Interkorrelationskoeffizienten bei Veränderung der Stichprobenvarianzen. Wir wollten untersuchen, ob die in der Literatur häufig behauptete Minderung von Korrelationskoeffizienten eine notwendige Folge der Reduktion der Varianz einer oder mehrerer zu korrelierender Variablen sei.

Im Kapitel 8 haben wir uns mit dem Phänomen der "Regression zur Mitte " bei Meßwiederholungen beschäftigt. Wir hatten uns die Frage gestellt, ob Zunahme oder Abnahme von Testleistungen bei Meßwiederholungen - wie in der Fachliteratur behauptet - stets Methodenartefakte sind, oder ob es sich dabei doch um echte, psychologisch relevante Ver-

Änderungen handle.

Im Kapitel 9 haben wir die Frage untersucht, ob mit Hilfe der multiplen Diskriminanzanalyse für Variablen des kognitiven Bereichs die individuelle Schullaufbahnberatung und Entscheidungsfindung verbessert werden könne.

Im letzten Kapitel des ersten Teils dieser Arbeit haben wir das Problem der Schätzung der Umweltvarianz in der Anlage-Umwelt-Kontroverse zur Intelligenzentwicklung aufgegriffen. Die umfangreichen Daten über die Schüler der Längsschnittstudie eröffnen die Möglichkeit, Pseudozwillingspaare zu bilden. Wir wollten herausfinden, wie hoch der Anteil der Umweltvarianz - gemessen durch die Korrelationen der Gesamtstandardwerte - dieser Paare sei. Zudem konnten wir die Varianzschätzung zu zwei verschiedenen Meßzeitpunkten berechnen, ein Aspekt, der in dieser Kontroverse unseres Wissens deswegen unberücksichtigt bleiben mußte, weil die dazu notwendigen Daten nicht zur Verfügung standen. Ergänzend dazu haben wir für einige Zwillingsgruppen aus der Fachliteratur homogenere Untergruppen gebildet und dafür neue Korrelationskoeffizienten berechnet. Die Homogenisierung sollte uns Aufschluß über das Zustandekommen der Korrelationskoeffizienten liefern. Von unseren Ergebnissen erwarten wir Anregungen für die Diskussion des herkömmlichen Modells der Varianzschätzung in der Anlage-Umwelt-Kontroverse.

Den zweiten Teil unserer Arbeit bildet eine Kontrolluntersuchung. Hauptziel war es, die Übereinstimmungsvalidität der PMA 13-14 der Längsschnittstudie zu ermitteln, d.h. diese Untersuchung wurde durchgeführt, um die PMA 13-14 mit geeigneten Tests, dem "Prüfsystem für Schul- und Bildungsberatung" (P-S-B) und dem "Frankfurter-Analogie-Test"(FAT 7-8), vergleichen zu können. Dabei haben wir die Gelegenheit genutzt, unsere Befunde aus der Längsschnittstudie hinsichtlich der Veränderungsmessung kognitiver Strukturen, sowie der Verwendung der multiplen Diskriminanzanalyse als Hilfsmittel bei der Schullaufbahnberatung einer ersten Überprüfung zu unterziehen.

In der Forschungspraxis werden Stichproben leider häufig so unvollständig beschrieben, daß Literaturvergleiche hinsichtlich der Ergebnisse verschiedener Untersuchungen fast beliebig und die Replizierbarkeit der Untersuchungen auf Grund der Angaben oft kaum möglich erscheinen. Ausgehend davon, dieser anhaltende Trend solle unterbrochen werden, haben wir daraus die Konsequenz gezogen, die eigenen Stichproben, insbesondere die Längsschnittstichprobe, so detailliert wie möglich zu beschreiben. Dazu haben wir die uns bekannten Merkmale der Probanden in einer mehrdimensionalen - geschachtelten - Weise dargestellt.

Wir meinen, daß diese an multifaktorielle Varianzanalysepläne angelehnte Darstellungsweise den sicheren Vorteil hat zu übersehen, wie die einzelnen Felder (Zellen) besetzt sind und ob daraus bereits spezifische - möglicherweise verzerrende - Effekte zu erwarten sind.

Die Stichprobe der Längsschnittuntersuchung wurde unter den folgenden Aspekten dargestellt:

- 1) Die Ausgangslage des gesamten Einschulungsjahrgangs im Jahre 1960/61

Die Stichprobe umfaßt den gesamten Einschulungsjahrgang an fünf Volksschulen (Grundschulen) in Stadtbezirken von Münster, insgesamt 341 Schüler/innen. Davon konnten 20 Pbn (= 6%) nicht erfaßt werden, sechs weitere Pbn wurden später wegen unvollständiger Angaben ausgeschieden. Die Stichprobenverteilung wurde nach den sozio-demographischen Merkmalen "Einschulungstermin", "Geschlecht" und "Alter" zu den drei Untersuchungsterminen zusammengestellt (vgl. Tabelle 1) und hinsichtlich der Merkmalsverteilung und möglicher Schrumpfungseffekte überprüft.

Ergebnisse:

- Die Stichprobenschrumpfungen von Untersuchung zu Untersuchung weisen keine bedeutsamen Effekte auf.

- Von der Zurückstellung zu Beginn der Schulpflicht sind Geschlecht und Alter jeweils signifikant verschieden betroffen: Am häufigsten jüngere Jungen, dann jüngere Mädchen, ältere Jungen und ältere Mädchen.

2) Einfluß der Grundschule auf die weiteren Bildungschancen

Nach der dritten Datenerhebung (1968) wurde retrospektiv die Frage untersucht, in welchem Maße die abgebenden Grundschulen die Bildungschancen bzw. den Bildungsweg der Schüler beeinflußt haben(könnten). Dazu wurden die Pbn nach den Merkmalen "besuchte Grundschule", "Alter", "Geschlecht", "Einschulungstermin" und zunächst besuchte "Schulart nach Abschluß der Grundschulzeit" gruppiert. Außerdem wurden die vor Verlassen der Grundschule aus der Stichprobe ausgeschiedenen Pbn nach den genannten Merkmalen gesondert hinzugefügt (vgl. Tabelle 2).

Ergebnisse:

- Verteilungsabweichungen hinsichtlich der Merkmale "Geschlecht" und "Alter" sind zufällig.
- Der Übergang zu einer RS oder OS ist sehr signifikant verschieden in den fünf Grundschulen.
- Je ein Drittel aller SS entstammen nur zweier GS.
- Der Stichprobenschwund von 1960/61 bis 1968 ist

in den fünf Grundschulen sehr bedeutsam verschieden. Von 235 Schülern konnte festgestellt werden, daß nach der Grundschulzeit 5% zur SS, 40% zur HS, 21% zur RS und 34% zur OS Übergewechselt sind. Die tatsächlichen Quoten von RS und OS dürften - bezogen auf N=315 - noch größer sein, da z.B. Mobilität und Bildungsaspiration positiv korrelieren (vgl. BOLTE/RECKER 1976).

- Die Einschulungs- bzw. Zurückstellungspraxis ist hochsignifikant verschieden in den Grundschulen.

3) Prüfung des Stichprobenschwundes mittels psychologischer Variablen

Um möglichst viele Informationen über die Veränderung der Ausgangsstichprobe von 1960/61 bis 1968 zu erhalten, wurden auch psychologische Variablen als Prüfkriterien hinzugezogen; hier: Schulreife- und Intelligenztestergebnisse (vgl. Tabelle 3). Bei insgesamt zehn solcher Vergleiche traten keine bedeutsamen Mittelwertunterschiede und lediglich in zwei Fällen signifikante Varianzunterschiede (F-Test) auf.

Aus den bisherigen Ergebnissen kann gefolgert werden, daß zwar die jeweils verbliebene Stichprobe noch repräsentativ für die Ausgangsstichprobe war, daß aber sozio-demographische Merkmale wie "Geschlecht",

"Alter" und "altersgemäße Einschulung" vs "Zurückstellung" den Geltungsbereich psychologischer Forschungsergebnisse, etwa vorgefundene Gruppenunterschiede oder Zusammenhänge vermutlich erheblich einschränken (können). Dies ist kein spezifisches Problem dieser Längsschnittuntersuchung, sondern es gilt - zumindest für Längsschnittstichproben - allgemein, wobei den genannten Klassifikationsmerkmalen auch nur Beispielcharakter zukommt. Zu selten wird in der Literatur eine auch nur den offensichtlich gegebenen Möglichkeiten nach präzise Beschreibung der untersuchten Gruppe(n) geliefert. Mangelnde Kontrolle möglicher Varianzquellen (vgl. FEGGER 1977) macht die Interpretation von Ergebnissen häufig nahezu beliebig. Davor schützt auch der Hinweis auf ähnliche oder gar gleiche Ergebnisse anderer Publikationen nicht.

4) Die Zusammensetzung der Stichprobe nach maximal
8 Jahren Schulbesuch im Jahre 1968

Die Längsschnittuntersuchung umfaßte 1968 noch N=224 Probanden, die nach den Merkmalen Einschulungstermin, Geschlecht, Alter, Schicht (Beruf des Vaters), besuchte Schulart, Zeitpunkt des Übergangs auf eine weiterführende Schule und (Nicht-)Versetzung klassifiziert wurden (vgl. Tabelle 4). Eine derartige Aufschlüsselung der Stichprobe soll auf die im allgemeinen in der Literatur selten in Betracht gezogenen

sozio-demographischen Veränderungen von Untersuchungsgruppen, die dann häufig daraus resultierenden signifikanten erwartungswidrigen Merkmalsverteilungen hinweisen und demzufolge auf die nicht geringe Wahrscheinlichkeit aufmerksam machen, wie leicht Ergebnisse erzielt werden können, die in Wirklichkeit jedoch nur Scheineffekte darstellen oder darstellen können, wenn oder weil sie auf Stichprobenfehler zurückgeführt werden können.

Ergebnisse:

In den folgenden Fällen ergaben sich zumindest signifikante ($p < 5\%$) Verteilungsabweichungen.

- Bei der Einschulung werden mehr jüngere Kinder zurückgestellt.
- Von der Zurückstellung sind Schüler/innen unserer beiden unteren Sozialschichten (Beruf des Vaters) häufiger betroffen.
- Elf von zwölf Sonderschülern gehören zu den Zurückgestellten.
- Die Chancen der Zurückgestellten, später eine RS oder OS zu besuchen sind stark reduziert. Kein einziger zurückgestellter Schüler aus der Gruppe "Älteres Halbjahr" gelangt auf eine weiterführende Schule.
- Unter den zurückgestellten RS und OS finden sich erwartungswidrig viele Repetenten (Sitzenbleiber).

Bis 1968 sind bereits 44% der Zurückgestellten, aber nur 24% der altersgemäß Eingeschulten mindestens einmal sitzengeblieben. Bis einschließlich Sommer 1967 haben bereits 51% der Jungen und Mädchen eine "Behinderung" in ihrer normalen Schullaufbahn infolge Zurückstellung, Sitzenbleiben, verspäteten Übergang zu einer weiterführenden Schule hinnehmen müssen. Jeder 8. Schüler ist bereits im Sommer 1967 um 2 Jahre Überaltert.

- Bezüglich der beiden Altershalbjahre gehören mehr Jüngere den beiden oberen Sozialschichten an, entsprechend weniger der Schicht III. Bei den Schichten IV und V sind die Verhältnisse ausgeglichen.
- Erwartungswidrig viele Kinder der beiden oberen Schichten wurden im Frühjahr geboren, das für die Entwicklung als besonders vorteilhaft bezeichnet wird.
- Es gehen tendenziell ($p < 10\%$) mehr Jüngere nach der fünften Klasse zu einer RS oder OS.
- Sowohl in den Sonderschulen als auch in den Oberschulen finden sich mehr Jungen als Mädchen. Dieses Mißverhältnis wird noch deutlicher, wenn man diesen Vergleich nur für die Jungen und Mädchen der RS und OS anstellt.
- Insgesamt gehen mehr Kinder bereits nach der vierten Klasse zur OS als zur RS. Jungen gehen häufiger als Mädchen bereits nach der vierten

Klasse zur OS.

- Unter dem Schichtaspekt (Beruf des Vaters) ist festzustellen, daß kein SS oder HS der oberen Schicht (I) und nur zwei SS der Schicht II entstammen, die restlichen neun SS zählen zur Schicht IV oder V. Acht Pbn der Schicht I gehen zur OS, einer zur RS. Insgesamt bestehen für die Jungen und Mädchen der Schicht IV und V - statistisch gesehen - erhebliche Bildungsbenachteiligungen.
- Unter dem Schichtaspekt scheint ferner interessant, daß eine Tendenz ($p < 10\%$) vorherrscht, ein Kind entweder möglichst früh oder gar nicht zur RS oder OS zu schicken.
- Für das Sitzenbleiben findet sich lediglich in der Gruppe der OS eine erwartungswidrige Verteilung. Pbn der Schicht III müssen hier häufiger eine Klasse wiederholen als diejenigen der übrigen Schichten.

In dem Kapitel über die instrumentellen Eigenschaften der Verfahren haben wir zunächst feststellen müssen, daß trotz großer Sorgfalt einige Tests nicht normal verteilt sind. Neben anderen Überlegungen gelangten wir zu der Überzeugung, eine repräsentative Stichprobe sei eine unverzichtbare Bedingung für die Eichung von Tests. Wir haben daraus die Frage abgeleitet, welche diesbezüglichen Informationen die

die Handanweisungen bekannter deutschsprachiger Intelligenztests dem Anwender bereitstellen und einige daraufhin untersucht. Die Ergebnisse waren nicht sehr ermutigend. Bei keinem der Tests ist der Anwender in der Lage, das Vorliegen der Eichvoraussetzungen an Hand mitgeteilter Informationen vollständig nachvollziehen zu können. Viel mehr wird er zuweilen eher mit der Frage konfrontiert, ob denn die Eintragungen der vorliegenden Normtabellen eine vollständige empirische Basis haben können. U.E. müßten hier erhebliche Anstrengungen unternommen werden, will man sich nicht - langfristig vermeidbar - der öffentlichen Kritik aussetzen, die durchaus auch hier zu juristischen Interventionen - wie in den USA bereits geschehen - führen kann, mit all den negativen Begleiterscheinungen, die eine Diskussion in der Öffentlichkeit unvermeidbar auslösen würde.

Im weiteren Verlauf dieses Kapitels haben wir unser Interesse auf die Interkorrelationen der Testwerte in Abhängigkeit von der Varianzentwicklung von Substichproben gelenkt. In Publikationen ist davon die Rede, daß die Korrelation zweier Merkmale sich reduziere, wenn die Stichprobe homogenisiert werde. Als Folge davon werden dann auch Korrelationskoeffizienten häufig minderungskorrigiert, obgleich nicht weiter nachgewiesen wird, daß die Voraussetzungen dafür

gegeben seien.

Wir glauben nachgewiesen zu haben, daß bei dieser Vorgehensweise ein wichtiger Aspekt übersehen wird. Wenn Stichproben homogenisiert werden, geschieht dies häufig über sozio-demographische Merkmale, d.h. die zu korrelierenden Merkmale bleiben dabei außer Betracht. Dies aber bedeutet, daß die Kovarianz als der relevanten Größe (nicht die Varianz) sich nicht notwendig überproportional reduzieren muß, gegenüber dem Produkt der Varianzen der Meßwertreihen. Nur wenn dies jedoch geschieht, reduziert sich auch die Korrelation. Theoretisch kann eine Korrelation bei konstanter Kovarianz und reduzierter Varianz sogar steigen, wie unsere Korrelationstabellen zeigen.

Wenn also eine bestimmte Korrelation X ihren Wert ändern soll, so muß sich für den Anstieg die Kovarianz relativ zum Produkt der Meßwertvarianzen weniger und für einen Abfall relativ mehr ändern. Es müßte in empirischen Untersuchungen ein zusätzlicher Akzent systematisch aufgenommen werden, nämlich die Frage, wodurch sich Korrelationskoeffizienten bei der Bildung von Substichproben im Einzelfall verändern. Als Methode dazu bietet sich u.a. die Kovarianzanalyse an. Zwei mögliche Hauptursachen wären dabei von besonderem Interesse, die Tests und die

Stichproben. Bezüglich der Tests könnten sich aus der angesprochenen Frage Hinweise für Testverbesserungen ergeben. Zum zweiten könnten Einsichten gewonnen werden, die eine größere Sorgfalt bei der Stichprobengewinnung angezeigt sein lassen. Beides zusammen könnte die Voraussetzung schaffen helfen, eine bessere Basis, als derzeit vorhanden, für Ergebnisvergleiche zu finden.

Zu den von uns ermittelten Korrelationskoeffizienten: Die Korrelation zwischen Gesamtstandardwerten beider PMA-Batterien beträgt 0,85. Auch die übrigen Koeffizienten analoger Tests sind hoch genug, um die Testergebnisse als Stabilitätsmaße verwenden zu können. Wenn man die Korrelationsmatrizen homogener Stichproben (SS, HS, RS, OS) miteinander vergleicht, so ergibt dies kein einheitliches Bild. Die Koeffizienten entsprechen vielfach nicht den auf Grund der Varianzänderung herkömmlicher Weise zu erwartenden Werten. Diese Ergebnisse zeigen deutlich den besonderen Stellenwert der Kovarianz. Wegen der Fülle der Korrelationskoeffizienten, deren Bedeutung erst in der Gegenüberstellung deutlich werden kann, wie dies in den Korrelationsmatrizen geschehen ist, verweisen wir den an dieser Stelle auf die Matrizen selbst.

Hinsichtlich der Validität der beiden Schulreife-
tests und der PMA-Batterien konnten hier den von
KEMMLER (1976) mitgeteilten einige wichtige
Koeffizienten hinzugefügt werden. So haben wir eine
Schulstatuskennziffer, die sich für jeden Schüler
aus der besuchten Schulart und Klasse zusammen-
setzt, mit den Schulreifetests und den "Skores"
Verbal, Reasoning, Lesefrei und GSW der Unter-

suchungen 1964 und 1968 korreliert. Die Korrelationen sind im großen und ganzen in der Höhe mit denen der Gesamtnoten 1964 und 1968 vergleichbar. Die Korrelation zwischen Schulstatus und GSW 9-10 beträgt 0,73 und mit dem GSW 13-14 0,71. Besonders deutlich erscheint die Stabilität des Faktors Verbal mit 0,71 bzw. 0,73. Außerdem ist die wachsende Bedeutung des Faktors Reasoning 1964 mit 0,57 und 1968 mit 0,70 nicht zu übersehen. Diese Ergebnisse werden noch durch die beträchtlich niedrigeren Korrelationen der Skores Lesefrei 1964 mit 0,66 und 1968 mit 0,48 unterstrichen.

Im folgenden Kapitel haben wir versucht, die Frage der Veränderungsmessung zu akzentuieren. Es ging darum herauszufinden, ob sich bei Schülern zwischen deren 10. und 14. Lebensjahr eine Veränderung kognitiver Strukturen vollziehe. Dabei sieht man sich dem Problem gegenüber, wie eine derartige Veränderung gemessen werden könne. Damit sind zwei wichtige Aspekte angesprochen, die Reliabilität von Messungen und das Phänomen der Regression zur Mitte. Die Schwierigkeit besteht darin, daß bei sehr reliablen Messungen Unterschiede zwischen erster und zweiter Messung per definitionem dem Meßfehler zuzurechnen sind. Tatsächliche Unterschiede sind demnach paradoxerweise nur mit wenig reliablen Meßinstrumenten nachweisbar. In diesem Falle weiß man aber nicht mehr, ob erste und zweite Untersuchung

noch denselben Gegenstand messen. Beim Phänomen der Regression handelt es sich um eine statistische Erwartung; Probanden, die bei der ersten Messung überdurchschnittliche Werte erzielt haben, kommen bei der Wiederholungsmessung im allgemeinen auf niedrigere Werte, mit der Tendenz zur Populationsmitte. Und Probanden mit niedrigeren Werten in der Messung verbessern sich bei der Meßwiederholung, ebenfalls zur Populationsmitte. Der gegenwärtige Diskussionsstand eröffnet noch keinen Ausweg aus diesen prinzipiellen Schwierigkeiten (vgl. CRONBACH + FURBY 1970). Hier wurde auf der Basis der Argumentation von HELMREICH (1977) die Differenz zwischen zweiter und erster Messung als Modell für Veränderungsmessungen verwendet. Mit Hilfe dieses Modells ist versucht worden, eine Differenzierung zwischen statistisch zu erwartenden und psychologisch relevanten Ergebnissen zu ermöglichen.

Die Gesamttichprobe wurde dazu in Untergruppen nach Schulart aufgeteilt und die Differenzen analoger Tests getrennt berechnet und zusammengestellt.

Ergebnisse:

- Die Veränderungsquoten zwischen GSW_{14} - GSW_{10} entsprechen für die einzelnen Schularten der statistischen Erwartung.

- Zwischen USW-Verbal₁₄ - USW-Verbal₁₀ Überwiegen bei den SS die negativen und bei den OS die positiven Veränderungen.
- Deutlich entgegen der statistischen Erwartung verhalten sich die Differenzverteilungen bezüglich USW-Reasoning₁₄ - USW-Reasoning₁₀. Während OS und RS zum Teil erhebliche positive Differenzen (im Vergleich zu den negativen) erzielen, Überwiegen bei den HS die negativen.
- Beim USW-Lesefrei₁₄ - USW-Lesefrei₁₀ sind die Verhältnisse den Reasoning-Befunden genau entgegengesetzt. Bei den OS findet sich die größte Quote negativer und bei den HS die größte Quote positiver Veränderungen.
- Bei den Differenzen zwischen den Schulreifetests und den GSW's 1964 und 1968 gibt es den durchgehenden Befund, daß die SS und HS jeweils erwartungswidrig große negative und die OS entsprechend erwartungswidrig positive Veränderungsquoten aufweisen. Die Differenzen mit dem KERN-Schulreifetest sind wegen mangelnder Gütekriterien nur schwer in ihrer Bedeutung einzuschätzen.

Eine Interpretation dieser Ergebnisse ist dann möglich, wenn die Mittelwerte, welche die einzelnen Gruppen (SS, HS, RS, OS) in den miteinander verglichenen Tests erreicht haben, unter dem Aspekt des

Phänomens der Regression zur Mitte einbezogen werden.

Obgleich die OS bei den Testvergleichen schon für die jeweils erste Messung Überdurchschnittliche (die höchsten) Mittelwerte erzielen, ergibt sich vielfach noch ein bedeutsamer Zuwachs.

Bei den SS und HS ist es genau umgekehrt; obgleich die Mittelwerte der ersten Messung immer unterdurchschnittlich sind, kann man mit Ausnahme der analogen Skores Lesefrei keinen Zuwachs, sondern tendenzielle oder signifikante Leistungsabnahmen konstatieren.

Diese Ergebnisse laufen dem Phänomen der Regression zur Mitte zuwider, denn danach hätten sich die Mittelwerte von HS und OS, bezogen auf bestimmte Tests, aufeinander zubewegen sollen. Wir folgern daraus, daß es neben dem statistischen Phänomen der Regression zur Mitte noch einen psychologisch relevanten Aspekt (echte Veränderungen) gibt, der es gestattet, diese unerwarteten Ergebnisse zu deuten.

Zwei Einflußgrößen, die sich aus der Gruppierung der Substichproben ergeben, könnten dabei von Bedeutung sein; die soziale Herkunft der Schüler und

die spezifischen Anforderungen, die die besuchte Schule an die Schüler stellen. Es wird hier für wahrscheinlich gehalten, daß der statistisch nicht zu erwartende Leistungszuwachs der OS (im Vergleich zu den HS) von 1960 bis 1964 auf besondere häusliche Anregungen der späteren OS zurückzuführen ist, denn die OS entstammen überwiegend unseren oberen Schichten, den Schichten mit besonderen Bildungsaspirationen.

Das Resultat, wonach die OS von 1964 bis 1968 signifikanten Zuwachs im Reasoning-Faktor und signifikante Abnahme im Skore Lesefrei erzielen (bei den HS ist dieser Sachverhalt genau umgekehrt), deuten wir so, daß Schlußfolgerndes Denken, wie es der Reasoning-Faktor messen soll, die Fähigkeit ist, welche die Schüler der Oberschulen in allen Unterrichtsfächern erlernen und verbessern, so daß es sich bei dem Reasoning-Zuwachs um eine echte Zunahme handelt. Praktische Intelligenz (Skore Lesefrei) scheint hier keine besonders große Rolle zu spielen.

Will man die Befunde der Hauptschüler deuten, so läßt sich vermuten, daß die HS in erster Linie Fähigkeiten der praktischen Intelligenz (Skore Lesefrei) verbessern. Hier scheinen Anforderungen des Schlußfolgernden Denkens weniger wichtig zu sein.

Sollten sich diese Unterschiede als allgemein vorfindbar erweisen, wäre die Frage zu stellen, ob damit für die HS die grundgesetzlich verankerte Entfaltungsgarantie eine unbillige Einengung erfahre.

Unter den Probanden haben wir sechzehn gefunden, deren individuelle kognitive Strukturen sich signifikant verändert haben; bezogen auf den GSW haben sich neun signifikant verschlechtert und sieben signifikant verbessert. Wegen der kleinen Zahl war eine Typisierung der Pbn nicht möglich. Die Regression der GSW zur Stichprobenmitte ist unverkennbar. Dieses Ergebnis muß also nicht unbedingt psychologisch bedeutsam sein, wenngleich die Unterschiede beider Gruppen in einigen nicht-kognitiven Variablen Hinweise liefern, dieses Problem genauer zu studieren (vgl. MANDL 1975, S. 181). So haben Absteiger weniger Angst, schätzen ihre Schulleistungen und ihre Schulanpassung geringer ein als die Aufsteiger. Ferner suchen die Absteiger die Ursachen für Mißerfolg eher bei sich selbst. Nach Ansicht ihrer Lehrer haben die Absteiger weniger Kontakt zu Schulkameraden und widersetzen sich öfter den Anordnungen ihrer Lehrer als die Aufsteiger.

Es sieht also so aus, als seien hier ein sta-

tistisches Phänomen (Regression zur Mitte) und ein psychologischer Problemkomplex kontaminiert. Die sehr wenigen u.E. aber interessanten Ergebnisse zeigen einen Weg, wie man das Aufsteiger/Absteiger-Problem dennoch untersuchen kann, auch wenn dabei die oben aufgezeigten prinzipiellen Schwierigkeiten nicht überwunden werden können; nämlich, indem Merkmale untersucht werden, die nicht als Kriterium für Ab- oder Aufsteiger gedient haben. In einem ersten Schritt könnten Ergebnisse zumindest unter dem Gesichtspunkt der Plausibilität gedeutet werden. Daraus könnte möglicherweise eine gewisse Sensibilität der Lehrer gegenüber bestimmten Schülern erwachsen.

Da immer noch nach brauchbaren Modellen zur Schullaufbahnberatung gesucht wird, wollten wir die Möglichkeit der Längsschnittstudie nutzen, retrospektiv zu ermitteln, ob die "Längsschnittkinder", die alle ohne Einsatz von Tests als Entscheidungshilfe ihre jeweilige Schullaufbahn getroffen und durchlaufen haben, in ihrer Mehrzahl die ihren kognitiven Strukturen entsprechende Wahl getroffen haben.

Die kognitive Strukturierung (Intelligenz) der Schüler verschiedener Schullaufbahnen nachzuzeichnen ist deshalb interessant und wichtig, weil

sie in der Bildungsdiskussion als das zentrale Kriterium dafür angesehen wird, welche Schullaufbahn jemand absolvieren sollte. Wir haben uns dazu der multiplen Diskriminanzanalyse bedient. Das Verfahren ist kurz folgendes: Die Testergebnisse der HS, RS, und OS werden multiplen Diskriminanzanalysen unterzogen, die zuvor für die drei Schularten berechnet wurden. Als Ergebnis wird für jeden Schüler gemäß seiner kognitiven Struktur die angemessendste Schulart ermittelt.

Als Ergebnis können wir feststellen, daß die multiple Diskriminanzanalyse ein sehr gutes Hilfsmittel zur Schullaufbahnberatung sein könnte. Wenn man - wie hier geschehen - die kognitive Struktur als Entscheidungskriterium verwendet, ist festzustellen, daß Chancengleichheit im Bildungswesen in unserer Stichprobe nicht voll realisiert ist. Es gibt eine Gruppe von 35 Pbn (=17%), die als HS eher eine RS oder OS und als RS eher eine OS hätten besuchen sollen. Daneben gibt es aber auch eine Gruppe von 39 (=18%) RS und OS, die - entsprechend ihrer kognitiven Struktur - eine sie "überfordernde" Schullaufbahn gewählt hat. Die Güte der Zuordnung wird mit Hilfe des tatsächlichen Schulerfolgs überprüft und kann als zufriedenstellend bezeichnet werden, obgleich einige der verwendeten Tests noch nicht die

gewünschte Güte aufweisen. Dazu kommt das Fehlen von motivationalen und anderen Merkmalen. Wenn also die Testmängel beseitigt und andere relevante Merkmale mit in die Analysen bzw. Beratungen einbezogen werden, können die Schullaufbahnberatungen offensichtlich beträchtlich verbessert werden.

Folgt man der Diskussion im Band 3 (1979) der Fachzeitschrift "INTELLIGENCE", so werden sich zumindest in den USA die Intelligenzforschung und der Einsatz von Tests zukünftig eher verstärken. Fraglich bleibt freilich, ob es die bereits bekannten Tests sein werden; außer Frage steht, daß die Intelligenztheorie noch aussteht (BROWN/FRENCH 1979; DETTERMANN 1979; HORN 1979; NEISSER 1979; RESNICK 1979; TURNBULL 1979).

Im letzten Kapitel des ersten Teils dieser Arbeit wurde versucht, die Bedeutung des Umwelteinflusses für die Intelligenzentwicklung abzuschätzen. Dabei haben wir uns - wie auch in der Literatur üblich - der Korrelationen zwischen Nichtverwandten bedient. Dazu wurden Pseudo-Zwillinge nach sozio-demographischer Ähnlichkeit gebildet. Mit den sehr begrenzten Möglichkeiten unserer Stichprobe konnten wir feststellen, daß die dann erhaltenen mittleren Korrelationen mit 0,50 erheblich über den in der

Literatur für derartige Schätzungen der Umweltvarianz liegenden angesiedelt sind. Zweitens konnten wir im Längsschnittaspekt feststellen, daß sich die Koeffizienten mit zunehmendem Alter erheblich reduzieren. Diesem Altersaspekt wird bei der Darstellung von Korrelationskoeffizienten u.E. nicht genügend Aufmerksamkeit geschenkt. Vielmehr werden in der Literatur häufig Zwillingsstichproben mit einer großen Altersspanne zusammengestellt, was die Korrelation beeinflußt. Das Gleiche gilt auch für das Geschlecht und die Schichtzugehörigkeit. Aus diesem Grunde haben wir die Zwillingsgruppen der Untersuchungen von NEWMAN/FREEMAN/HOLZINGER (1973) und SHIELDS (1962) homogenisiert und neue Korrelationskoeffizienten berechnet. Die Differenzen zu den von den Untersuchern mitgeteilten Koeffizienten sind so groß, daß empfohlen werden kann, bei zukünftigen Untersuchungen "fremde" Varianz aus den Korrelationskoeffizienten herauszupartialisieren.

Es versteht sich von selbst, daß unser Versuch der Varianzschätzung nicht für sich beanspruchen kann, der "Wahrheit" näher gekommen zu sein. Dazu ist das Datenmaterial nicht umfangreich genug. Sollten sich die Ergebnisse aber mit größeren Datenmengen replizieren lassen und zu ähnlichen Ergebnissen führen,

könnte dies zur Folge haben, daß die Zwillingsforschung als Modell zur Varianzschätzung in der Anlage-Umwelt-Kontroverse Überdacht werden müßte.

Im zweiten Teil dieser Arbeit ging es in erster Linie darum, mit Hilfe einer Querschnittuntersuchung die Validität der PMA 13-14 zu prüfen. Dazu wurde die Übereinstimmungsvalidität der PMA 13-14 mit dem PSB von HORN ermittelt. Die Korrelation zwischen GSW 13-14 und PSB-G1 beträgt 0,74, liegt damit im mittleren Bereich der Erwartungen. Der ebenfalls in diese Validierung einbezogene FAT 7-8 korreliert mit dem analogen Skore der PMA, USW-Reasoning 13-14, 0,68. Die nur mittlere Höhe dieser Korrelationskoeffizienten macht es ein bißchen schwierig, die Tests jeweils als Paralleltests zu bezeichnen, die sie ja nach den jeweiligen Modellannahmen für ihre Konstruktion eigentlich sein sollten. MANDL (1975, S. 73) fand bei Schülern des 4. Schuljahres zwischen PSB und PMA eine Korrelation von 0,70. Wegen der verschiedenen Stichprobe MANDL's und dieser Querschnittuntersuchung sowie der unterschiedlichen Untersuchungsintervalle ist ein direkter Vergleich der Koeffizienten wenig sinnvoll.

Die durchgeführte Faktorenanalyse weist dem gegenüber eine zufriedenstellende Ähnlichkeit der Faktorenstruktur beider Testbatterien aus.

Auch die Regressionsgleichung für die beiden Gesamtstandardwerte erlauben zufriedenstellende gegenseitige Vorhersagen.

Nimmt man die Ergebnisse zusammen, so liegt die Übereinstimmungsvalidität beider Testbatterien im Rahmen dessen, was in der Literatur berichtet wird.

Im diesem zweiten Teil der Arbeit haben wir auch versucht, die Frage der Veränderungsmessung nochmals aufzunehmen. Die Schüler erzielten im PSB eine sehr bedeutsam höhere Gesamtleistung als im PMA. Darüber hinaus gibt es den größten "Zuwachs" im Vergleich der Reasoning-Skores, also zwischen FAT und PSB-Reasoning einerseits und PMA-Reasoning andererseits. Der "Zuwachs" findet sich auf seiten der in der Abfolge zuletzt durchgeführten Tests. Dieser Befund kann mit Hilfe eines Lerneffekts erklärt werden, muß also kein echter Zuwachs sein. Es stellt sich aber die Frage, warum auch hier gerade in den Reasoning-Skores Veränderungen stattfinden und warum diese Veränderungen die größten Werte erreichen. Denkbar ist immerhin, daß eine Ähnlichkeit der Aufgaben der Testbatterien hier einen besonders großen Lerntransfer ermöglicht bzw. erfordert. Es bedarf allerdings gezielt angelegter Untersuchungen, um die tatsächlichen Ursachen herauszufinden.

Ähnlich wie beim höheren Gesamtstandardwert des PSB im Vergleich zum PMA für die Gesamtstichprobe lassen sich auch die um mindestens einhalb Standardwerteinheiten (15 Punkte) höheren Werte von 12 Probanden in 10 von 12 Fällen als Regressionseffekt deuten. Darauf deutet auch hin, daß die großen Differenzen nur von HS und RS erzielt werden. Die beträchtlichen Differenzen machen deutlich, wie groß die Möglichkeiten sind, auf Grund eines einzigen Wertes, wie der GSW ihn darstellt, Fehlurteile im Zuge einer Beratung zu fällen. Es drängt sich geradezu auf, nochmals auf die Vorzüge der Profildarstellung und -analyse zu verweisen und - wo möglich - mehrere Tests zum Einsatz zu bringen. Nicht zu erklären ist aus dem Datenmaterial warum die positiven Differenzen nur von Mädchen erzielt werden.

Mit Hilfe der multiplen Diskriminanzanalyse wollten wir überprüfen, ob sich das Ergebnis der Längsschnittstudie annähernd replizieren lasse und welche der beiden Testbatterien - systemimmanent - die genaueren Zuordnungen gestatte und ob ferner die systemimmanente Zuordnung erhöht werde, wenn beide Testbatterien gleichzeitig in die multiple Diskriminanzanalyse einbezogen würden.

Zum Ersten: Die Diskriminanzanalyse hat auch bei der Querschnittuntersuchung eine deutliche Diskriminanz der kognitiven Strukturen von Schülern verschiedener Schularten ergeben. Diese Untersuchung kann somit als Replikation der Längsschnittstudie gelten.

Zweitens: Mit Hilfe der PMA 13-14 gelingt die systemimmanente Zuordnung der Schüler besser als mit dem PSB.

Drittens: Eine gleichzeitige Einbeziehung beider Testbatterien erhöht die systemimmanente Zuordnung beträchtlich.

Des Weiteren wird deutlich, daß es eine nicht ganz unvermutet große Zahl von Probanden gibt, die gemäß ihrer kognitiven Struktur von der besuchten Schulart entweder unter- oder überfordert wird. Allerdings wird auch deutlich, daß die Quoten abhängig sind von den verwendeten Tests und deren Güte. Damit aber stehen unsere Befunde im Widerspruch zu denen TENT's (1969, S. 135), der feststellt, nicht eine Verbesserung der Beratungs- (Prognose-) instrumente sei anzustreben, sondern die Lösung der Schullaufbahnproblematik liege in einer Veränderung des Schulsystems.

Unsere Ergebnisse zeigen durchaus die Möglichkeit

und Notwendigkeit, die Tests, sofern sie bei der Schullaufbahnberatung eingesetzt werden sollen, erheblich zu verbessern. Dabei muß die Frage offenbleiben, ob auf der Basis stochastischer Testbatterien eine größere Effizienz zu erlangen ist, als dies die klassische Testtheorie erlaubt. Und was das Schulsystem selbst betrifft, so kann (nur) vermutet werden, daß sich für manche Schüler bessere Chancen eröffnen, wenn die Entscheidung, welche Schullaufbahn nach der Grundschule absolviert werden solle, nicht schon mit 10 oder 11 Jahren getroffen werden müßte.

Ein Fazit unserer Ergebnisse und Ausblick müßte lauten:

- Der detaillierten Darstellung der Forschungsdesigns müßte mehr Raum gegeben werden.
- Es müßte selbstverständlich werden, Unzulänglichkeiten eines Forschungsprojekts explizit machen zu können. Gerade diese Forderung mag als naiv betrachtet werden, da Publikationen ja nicht nur der Beförderung der Wissenschaft dienen. Wenn man jedoch bedenkt, was sicher nicht selten aus "Fehlern" zu lernen ist, sollte diese Forderung nicht leichtfertig übergangen werden.

Die Tätigkeit der Wissenschaftler soll - wenigstens

wird das in den letzten Jahren zunehmend gefordert -
(gesellschaftlich) relevanten Fragestellungen
gelten. Welche Relevanz-Position man auch bevor-
zugt, in einem ist HECKHAUSEN (1976, S. 3) zuzu-
stimmen, wenn er feststellt,

"Relevanz ist ein ex-post-facto-Kriterium. Man weiß
es gewöhnlich erst im nachhinein, ob es nicht bloße
Augenschein-Relevanz war."

- Eine Verbesserung der Tests und der Einsatz bisher
eher vernachlässigter statistischer Verfahren
zur Entscheidungshilfe bei Fragen der Schullauf-
bahnberatung können die Realisierung von Chancen-
gleichheit im Bildungswesen fördern helfen.

LITERATURVERZEICHNIS

- AMERICAN PSYCHOLOGICAL ASSOCIATION Standards for Educational and Psychological Tests. Washington D.C. 1974
- AMTHAUER, R.: Intelligenz-Struktur-Test I-S-T 70. Göttingen ⁴1973
- AMTHAUER, R.: Über die Prüfung der Zuverlässigkeit von Tests - erörtert am Intelligenz-Struktur-Test. Psychologische Rundschau, 1957, 8, 165-171
- ANASTASI, A.: Psychological Testing. New York ⁴1976
- ANASTASI, A. (Ed.): Testing Problems in Perspective. Washington D.C. 1966
- ANASTASI, A.: Vererbung, Umwelt und die Frage "Wie?" in SKOWRONEK, H. (Hrsg.): Umwelt und Begabung. Stuttgart 1973, S. 9-26
- AURIN, K.: Gleiche Chancen im Bildungsgang - Bericht der Bildungsberatungsstellen von Baden-Württemberg über Begabung und Schulleistung. Villingen 1968
- BAUMANN, U.: Psychologische Taxometrie. Bern 1971
- BELSER, H.: Testentwicklung. Weinheim ³1975
- BELSER, H./ANGER, H./BARGMANN, R.: Frankfurter Analogietest FAT 7-8. Weinheim 1965
- BETHÄUSER, H./REICHENBECHER, H.: Approximative Expertenabstimmung in der Bildungsberatung. in HELLER, K. (Hrsg.): Handbuch der Bildungsberatung, Bd. III, Stuttgart 1976, 861-877
- BLACK, F.W.: Season of Birth and Intelligence in a Sample of Learning-Disabled Children. Journal of Genetic Psychology, 1973, 123, 31-34
- BLOOM, B.: Stabilität und Veränderung menschlicher Merkmale. Weinheim 1971
- BOLTE, K.M.: Sozialer Aufstieg und Abstieg. Eine Untersuchung über Berufsprestige und Berufsmobilität. Stuttgart 1959

- BOLTE, K.M./RECKER, H.: Vertikale Mobilität. in
KÖNIG, R. (Hrsg.): Handbuch der empirischen
Sozialforschung. Stuttgart 1976, Bd. 5, 40-103
- BONDY, C. (Hrsg.): Hamburg-Wechsler Intelligenz-
test für Kinder (HAWIK) - Handbuch - Bern ³ 1966
- BOWLBY, J.: Mütterliche Zuwendung und geistige Ge-
sundheit. München 1973
- BRICKENKAMP, R.: The Development of Psychological
Tests in German-speaking Countries. The German
Journal of Psychology, 1981, 5, 133-148
- BROWN, A.L./FRENCH, L.A.: The Zone of Potential
Development: Implications for Intelligence Test-
ing in the Year 2000. Intelligence, 1979, 3,
255-273
- BURT, C.: The Genetic Determination of Differences in
Intelligence: A Study of Monocygotic Twins
Reared Together and Apart. British Journal of
Psychology, 1966, 57, 137-153
- BURKS, B.S.: The Relative Influence of Nature and
Nurture upon Mental Development; A Comparative
Study of Foster Parent - Foster Child Re-
semblance. The 27th Year Book, 1928, 221-310
- BUSE, L.: Intraindividuelle Merkmalsvariation und
Validität eines Extraversionsfragebogens. Eine
Untersuchung zur Gültigkeit des Eigenschafts-
begriffs. Zeitschrift für Differentielle und
Diagnostische Psychologie, 1980, 1, 35-42
- CHAUDHURY, R.H.: Socio-Economic and Seasonal
Variations in Births: A Replication. Social
Biology, 1972, 19, 65-68
- CRONBACH, L.J./FURBY, L.: How We Should Measure
"Change" - or Should We? Psychological Bulletin,
1970, 74, 68-80 and 218
- DALEN, P.: Season of Birth. A Study of Schizophrenia
and Other Mental Disorders. Amsterdam 1975

- DANNEEL, R.: Lebenserwartung und Jahreszeit der Geburt. Aktuelle Gerontologie, 1974, 4, 49-51
- DER BUNDESMINISTER FÜR BILDUNG UND WISSENSCHAFT
Bildungsbericht '70. Bonn 1970
- DER SPIEGEL vom 17.12.1979
- DETERMANN, D.K.: A Job Half Done: The Road to Intelligence Testing in the Year 2000.
Intelligence, 1979, 3, 295-306
- DEUTSCHER BILDUNGSRAT Strukturplan für das Bildungswesen. Stuttgart ³1970
- DOORNBOS, K.: Age-Group Position and Scholastic Performance. Sociologica Neerlandica 1974, 10, 184-193
- DOORNBOS, K.: Geboortemaand en Schoolsucces.
Groningen 1971
- DORSCH, F.: Psychologisches Wörterbuch. Bern ⁹1976
- DÜKER, H./LIENERT, G.A.: Konzentrations-Leistungstest (K-L-T). Göttingen ²1965
- EBEL, R.L.: The Social Consequences of Educational Testing. in ANASTASI, A. (Ed.): Testing Problems in Perspective. Washington D.C. 1966
- ECKEL, K.: Der Korrelationskoeffizient bei verklumpten Variablen. Grundlagenstudien aus Kybernetik und Geisteswissenschaft, 1970, 11, 13-21
- ECKEL, K.: Die Bedeutung des Klasseneffekts für die schulpädagogische Forschung. Programmirtes Lernen und programmierter Unterricht, 1969, 6, 97-113
- ERLENMEYER-KIMLING, L./JARVIK, L.F.: Genetics and Intelligence: A Review. Science, 1963, 142, 1477-1479
- EYSENCK, H.J.: Psychology Is About People. London 1972
- EYSENCK, H.J.: Race, Intelligence and Education.
London 1971
- EYSENCK, H.J.: The Inequality of Man. London 1973,
dt. 1975

- FABER, E./NOLLAU, W.: Über einen Algorithmus zur mehrdimensionalen Diskriminanz-Analyse. Darmstadt 1969b
- FABER, E./NOLLAU, W.: Über ein Verfahren zur automatischen Klassifikation. Darmstadt 1969
- FEGER, H.: Zur Lage der Psychologie. Psychologische Rundschau, 1977, 28, 1-21
- FIEGENBAUM, W.: Abschlußuntersuchung einer Längsschnittstudie von 1960-1974 zum Schulerfolg. Untersuchung einer Stichprobe von Oberschülern mit dem Intelligenz-Struktur-Test (I-S-T 70) und einem Fragebogen zu Berufs- und Studieninteressen sowie zum schulischen Befinden. Münster 1974, unveröffentlicht
- FISCHER, G.H.: Abriss der klassischen Testtheorie. in FISCHER, G.H. (Hrsg.): Psychologische Testtheorie. Bern 1968, S. 15-53
- FUHRMANN, W.: Taschenbuch der allgemeinen und klinischen Humangenetik. Stuttgart 1965
- GADENNE, V.: Die Gültigkeit psychologischer Untersuchungen. Stuttgart 1976
- GNIECH, G.: Störeffekte in psychologischen Experimenten. Stuttgart 1976
- GREIF, S.: Gruppenintelligenztests - Untersuchungen am WIT, IST, LPS und ALT. Frankfurt 1972
- GROMBACH, H.H.: Konstanz und Variabilität von Persönlichkeitsmerkmalen - Probleme und Ergebnisse aus der Bonner gerontologischen Längsschnittstudie. Bonn 1975 (Phil. Diss.)
- HÄRNQUIST, K.: Relative Changes in Intelligence from 13 to 18. Scandinavian Journal of Psychology, 1968, 9, 50-64 and 65-82
- HARRIS, C.W. (Ed.): Problems in Measuring Change. Madison, Milwaukee ²1967
- HAVLICK, L.L./PETERSON, N.L.: Effect of the Violation of Assumptions Upon Significance Levels

- of the Pearson r. Psychological Bulletin, 1977, 84, 373-377
- HAYS, W.L.: Statistics for the Social Sciences. London ² 1974
- HECKHAUSEN, H.: Hoffnung und Furcht in der Leistungsmotivation. Meisenheim 1963
- HECKHAUSEN, H.: Relevanz der Psychologie als Austausch zwischen naiver und wissenschaftlicher Verhaltenstheorie. Psychologische Rundschau, 1976, 27, 1-11
- HECKHAUSEN, H./KEMMLER, L.: Entstehungsbedingungen der kindlichen Selbständigkeit. Der Einfluß der mütterlichen Selbständigkeit auf die seelisch-soziale Schulreife der Söhne. Zeitschrift für experimentelle und angewandte Psychologie, 1957, 4, 603-622
- HELLER, K.: Aktivierung der Bildungsreserven. Bern, Stuttgart 1970
- HELLER, K. (Hrsg.): Handbuch der Bildungsberatung. Bd. III, Methoden der Bildungsberatung und Bildungsforschung. Stuttgart 1976
- HELLER, K./NICKEL, H./ROSEMAN, B.: Beurteilen und Beraten. Stuttgart 1980
- HELLER, K.A./ROSEMAN, B./STEFFENS, K.-H.: Prognose des Schulerfolgs. Eine Längsschnittstudie zur Schullaufbahnberatung. Weinheim 1978
- HELLER, K./ROSEMAN, B.: Planung und Auswertung empirischer Untersuchungen. Stuttgart 1974
- HELLMUTH, J. (Ed.): Disadvantaged Child. Compensatory Education - A National Debate. New York 1970, Vol. 3
- HELMREICH, R.: Strategien zur Auswertung von Längsschnittdaten. Stuttgart 1977
- HERBIG, M.: Aufgabentypen zur Leistungsüberprüfung. in KLAUER/FRICKE/HERBIG/RUPPRECHT/SCHOTT: Lehrzielorientierte Tests. Düsseldorf 1972, S. 74-100

- HERRNSTEIN, R.J.: I.Q. in the Meritocracy. Boston-Toronto 1971, dt. 1974
- HILTMANN, H.: Kompendium der psychodiagnostischen Tests. Bern ³1977
- HINDLEY, C.B.: The Place of Longitudinal Methods in the Study of Development. in MÜNKES, F.J./HARTUP, W.W./WIT de, J. (Ed.): Determinants of Behavioral Development. New York 1972, 23-50
- HOFSTÄTTER, P.R.: Differentielle Psychologie. Stuttgart 1971
- HOJAK, J.: Untersuchungen zur Schulreife an zurückgestellten Schulanfängern. Münster 1962, unveröffentlicht
- HOLM, K. (Hrsg.): Die Befragung 6. München 1979
- HOLZKAMP, K.: Kritische Psychologie. Frankfurt 1972
- HOPF, D.: Möglichkeiten und Grenzen der Anwendung von Tests. in HOFER, M./WEINERT, F.E. (Hrsg.): Pädagogische Psychologie 2 - Lernen und Instruktion - Frankfurt 1973, S. 302-312
- HOPF/THIEL/REININGHAUS: Diktate aus dem deutschen Schrifttum. Frankfurt o.J.
- HORN, J.L.: Trends in the Measurement of Intelligence. Intelligence, 1979, 3, 229-239
- HORN, W.: Leistungsprüfsystem (L-P-S). Göttingen 1962
- HORN, W.: L-P-S Leistungsprüfsystem. Handanweisung für die Durchführung, Auswertung und Interpretation. Göttingen 1962
- HORN, W.: P-S-B Prüfsystem für Schul- und Bildungsberatung. Handanweisung für die Durchführung, Auswertung und Interpretation. Göttingen 1969
- HUSEN, T.: Begabung und Bildungspolitik. Die Bedeutung von Erbanlagen und Milieueinflüssen für die Bildungsreform. Hannover 1975
- HUSEN, T.: Der Einfluß der Schulbildung auf den IQ. in SÜLLWOLD, F.: Begabung und Leistung. Hamburg 1976, S. 241-262

- HUTHER, A.: Über das Problem einer psychologischen und pädagogischen Theorie der intellektuellen Begabung. Sammlung von Abhandlungen zur Psychologischen Pädagogik, 1910, 2, 397-437
- IM, H.-J.: Die Entwicklung eines europäischen Schlüsselwortes: Intelligenz und seine Bedeutung in der Wissenschaftssprache. Bonn 1975 (Phil. Diss.)
- INGENKAMP, K. (Hrsg.): Die Fragwürdigkeit der Zensurengebung. Weinheim 1971
- JÄGER, A.O. et al.: Vorwort. Diagnostica, 1979, 25
- JAMES, W.H.: Social Class and Season of Birth. Journal of Biosocial Science, 1971, 3, 309-320
- JANSSEN, J.P.: Studenten: die typischen Versuchspersonen psychologischer Experimente - Gedanken zur Forschungspraxis. Psychologische Rundschau, 1979, 30, 99-109
- JENCKS, C.: Chancengleichheit. Reinbek 1973
- JENSEN, A.R.: Bias in Mental Testing. London 1980
- JENSEN, A.R.: Genetics and Education. London 1972
- JENSEN, A.R.: How Much Can We Boost IQ and Scholastic Achievement? Harvard Educational Review, Reprint Series No. 2, ⁴1971, 1-123
- JENSEN, A.R.: Reducing the Heredity-Environment Uncertainty. Harvard Educational Review, Reprint Series No. 2, ⁴1971, 202-243
- KALLENBACH, K.: Der Einfluß unterschiedlicher Testanweisungen (Testfragen) auf die Testleistungen im HAWIK-Teilttest "Gemeinsamkeiten finden". Diagnostica, 1970, 16, 164-171
- KALVERAM, K.T.: Kompensatorische Kovarianz als Beispiel für einen Selektionseffekt. Archiv für die Gesamt Psychologie, 1969, 121, 255-265
- KANEKAR, S./MUKERJEE, S.: Intelligence, Extraversion, and Neuroticism in Relation to Season of Birth. The Journal of Social Psychology, 1972, 86, 309-310

- KEMMLER, L.: Eichungsfehler bei Tests. in KEMMLER, L./HECKHAUSEN, H. (Hrsg.): Praktische Fragen der Begabungsdiagnostik in der Erziehungsberatung. Weinheim 1965, S. 49-51
- KEMMLER, L.: Längsschnittuntersuchung vom ersten bis neunten Schuljahr einer Gruppe 1960 eingeschulter Kinder. Münster 1968, unveröffentlicht
- KEMMLER, L./LANGHEINRICH, D.: Eine deutsche Fassung der "Primary Mental Abilities" für acht- bis zehnjährige Kinder. Bericht über den 25. Kongreß der DGfP. Göttingen 1967, 669-673
- KEMMLER, L./LANGHEINRICH, D.: Ein Wissenstest für Gruppenuntersuchungen an acht- bis zehnjährigen Kindern. Diagnostica, 1969, 15, 177-185
- KEMMLER, L.: Schulerfolg und Schulversagen. Eine Längsschnittuntersuchung vom ersten bis zum fünfzehnten Schulbesuchsjahr. Göttingen 1976
- KEMMLER, L./HECKHAUSEN, H.: Ist die sogenannte "Schulreife" ein Reifungsproblem? in INGENKAMP, K. (Hrsg.): Praktische Erfahrungen mit Schulreife-tests. Basel 1962
- KEMMLER, L./KIEKHEBEN, F.: Vergleich von Gruppen-schulreifeverfahren. Schule und Psychologie, 1962, 9, 65-73
- KERN, A.: Der Grundleistungstest zur Ermittlung der Schulreife. Weinheim 1957
- KIEKHEBEN, F.: Vergleich von Gruppen-Schulreifetests. Der "Grundleistungstest", der "Göppinger Schulreifetest", der "Auslesetest für Schulneulinge" (Form B). Münster 1960, unveröffentlicht
- KIENER, F.: Jahreszeit der Geburt, Begabung und Schulleistung. Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie 1969, 1, 272-284
- KLAUER, K.J.: Die Bedeutung des Klasseneffekts für die schulpädagogische Furchung. Programmiertes Lernen und programmierter Unterricht, 1970, 7, 149-156

- KLAUER, K.J.: Das Experiment in der pädagogischen Forschung. Düsseldorf 1973
- KLEINER, A.: Göppinger Schulreifetest. Göppingen 38¹1965
- KNOCHE, W.: Das Erbe-Umwelt-Problem. Weinheim 1977
- KORNADT, H.-J.: Thematische Apperzeptionsverfahren. in HEISS, R. (Hrsg.): Handbuch der Psychologie, Bd. 6, 635-684
- KRAUTH, J./LINERT, G.A.: Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin. Freiburg/München 1973
- KREYSZIG, E.: Statistische Methoden und ihre Anwendungen. Göttingen 3¹1968
- KÜHN, R./HESSE, H.G.: Veränderungsmessung. Bericht über das Seminar 19 des IPN im Oktober 1979 in Kiel. in Deutsches Institut für Internationale Pädagogische Forschung - Mitteilungen und Nachrichten, Frankfurt 1979, Nr. 96/97, S. 119-124
- LABOUIE, E.W.: Measurement of Individual Differences in Intraindividual Changes. Psychological Bulletin, 1980, 88, 54-59
- LANGER, I./SCHULZ v. THUN, F.: Messung komplexer Merkmale in Psychologie und Pädagogik. München 1974
- LANGEVELD, M.J.: Über die Verwendung exakter Methoden bei der Auswertung von psychologischen Untersuchungen, Zeitschrift für Pädagogik, 1970, 16, 295-304
- LANGNER, H.-J.: Neustandardisierung der "Primary Mental Abilities" von Thurstone für 9-10jährige deutsche Kinder. Münster 1965, unveröffentlicht
- LEAHY, A.M.: Nature - Nurture and Intelligence. Genetic Psychology Monographs, 1935, 17, S. 237-308
- LENZ, W.: Medizinische Genetik. Stuttgart 1970
- LIEDEL, M.: Zurückstellung oder Vorschule, München 1972
- LIENERT, G.A.: Testaufbau und Testanalyse. Weinheim 2¹1967

- LUCHINS, A.S.: Mechanization in Problem Solving.
The Effect of Einstellung. Psychological
Monographs, 1942, 54
- LYKKEN, D.T.: Statistical Significance in Psychological
Research. Psychological Bulletin, 1968, 70,
151-159
- MACHADO, L.A.: The Development of Intelligence:
Politics or Science? Caracas 1981b
- MACHADO, L.A.: The Development of Intelligence -
A Political Outlook. Intelligence, 1981a, 5,
2-4
- MACHEMER, P.: Zum Problem der Legasthenie an der
Oberschule - Auslese und Verhaltenstherapeutische
Behandlung von Legasthenikern. Schule und Psycho-
logie, 1972, 18, 88-98
- MAGNUSSON, D.: Testtheorie. Wien 1969
- MANDL, H.: Kognitive Entwicklungsverläufe von Grund-
schülern. München 1975
- MCALL, R.B.: Challenges to a Science of Developmental
Psychology. Child Development, 1977, 48, 333-344
- MARTINDALE, C./BLACK, F.W.: Season of Birth and
Intelligence, Journal of Genetic Psychology,
1970, 117, 137-138
- MAST, M.: Bearbeitung eines neuentwickelten Interessen-
Milieu-Fragebogens. Münster 1969, unveröffentlicht
- MAYO, J./WHITE, O./EYSENCK, H.J.: An Empirical Study
of the Relation Between Astrological Factors
and Personality. Journal of Social Psychology,
1978, 105, 229-236
- MEILI, R./STEINGRÜBER, H.-J.: Lehrbuch der psycholo-
gischen Diagnostik. Bern ⁶1978
- MEILI, R.: Was ist und was will die psychologische
Diagnostik? Panel-Diskussion. in PULVER, U./
LANG, A./SCHMID, F.W. (Hrsg.): Ist Psychodiag-
nostik verantwortlich? Bern 1978, 186-229
- MELTON, A.W.: Editorial. Journal of Experimental
Psychology, 1962, 64, 553-557

- MERZ, F./STELZ, I.: Einführung in die Erbpsychologie. Stuttgart 1977
- MESCHOWSKY, H.: METHEMATIK als Grundlage. Plädoyer für ein rationales Bildungskonzept. München 1973
- METZGER, W.: Das Experiment in der Psychologie. Studium Generale, 1952, 5, 142-163
- MEYER, W.U.: Anspruchsniveau und erlebte Selbstverantwortlichkeit für Erfolg und Mißerfolg, Psychologische Beiträge, 1969, 11, 328-348
- MICHEL, L./TROST, G.: Vielleicht taugt der Hochschuleingangstest für Mediziner doch. Psychologie heute, 1980, Heft 1, S. 66-69
- MÜLLER, W./MAYER, K.U.: Chancengleichheit durch Bildung? Untersuchungen über den Zusammenhang von Ausbildungsabschlüssen und Berufsstatus. Stuttgart 1976
- NEISSER, U.: The Concept of Intelligence. Intelligence, 1979, 3, 217-227
- NEWMAN, H.H./FREEMAN, F.N./HOLZINGER, K.J.: Twins - Study of Heredity and Environment. Chicago 1937
- NIEMEIER, F.: Validierung eines Fragebogens zur Untersuchung der Selbstbeurteilung von Kindern der Grundschule. Münster 1968, unveröffentlicht
- NORDEN, I.: Anleitung zur Intelligenzprüfung nach Binet-Boberth "Binetarium". Göttingen 1953
- OTIS, A.S./LENNON, R.T.: Otis-Lennon Mental Ability Test. Technical Handbook. New York, Chicago, San Francisco, Atlanta, Dallas 1969
- OVERTON, W.F.: Concepts of Stability and Change. Human Development, 1978, 21, 360-363
- PASAMANICK, B./DINITZ, S./KNOBLOCK, H.: Socio-Economic and Seasonal Variations in Birth Rates. Milbank Memorial Fund Quarterly, 1960, 38, 248-254
- PAYNE, D.A./McMORRIS, R.F.: Educational and Psychological Measurement - Contributions to Theory and Practice, New York ²1975

- PELLEGRINI, R.J.: Birthdate Psychology: A new Look at Some Old Data. Journal of Psychology, 1975, 89, 261-265
- PETERMANN, F.: Veränderungsmessung. Stuttgart 1978
- RAATZ, U.: Neuere Ansätze zur Theorie der Reliabilität. in FISCHER, G.H. (Hrsg.): Psychologische Testtheorie. Bern 1968, 159-175
- REICHEL, G.: Analyse eines in Anlehnung an Thurstone's "PMA 11-17" konstruierten Gruppenintelligenztests für 13-14jährige. Münster 1969, unveröffentlicht
- REMBERG, A.: Differentialdiagnostische Untersuchungen an nicht schulreifen Lernanfängern. Münster 1961, unveröffentlicht
- RENN, H.: Zur Methodik der Verlaufanalyse: Das Problem der Nicht-Orthogonalität von Ausgangswerten und Faktoren der Änderung. Psychologische Beiträge, 1974, 16, 61-67
- RESNICK, L.B.: The Future of IQ Testing in Education. Intelligence, 1979, 3, 241-253
- REXILIUS, G.: Zur Problematik der Intelligenz und ihre Messung. in BRUDER, K.-J. et al.: Kritik der pädagogischen Psychologie. Falsche Theorien einer falschen Praxis. Reinbek 1976, 181-215
- RIEGEL, R.M./RIEGEL, K.F.: Standardisierung des Hamburg-Wechsler-Intelligenztests (HAWIE) für die Altersstufen über 50 Jahre. Diagnostica, 1959, 5, 97-128
- ROGAN, J.C./KESELMAN, H.J./MENDOZA, J.L.: Analysis of Repeated Measurement. The British Journal of Mathematical and Statistical Psychology, 1979, 32, 269-286
- ROSEMANN, B.: Prognosemodelle in der Schullaufbahn-Beratung. München 1978

- ROSSOLINO, G.: Die psychologischen Profile. Zur Methodik der quantitativen Untersuchung der psychischen Vorgänge in normalen und pathologischen Fällen. Eine experimental-psychologische Skizze. Klinik für psychische und nervöse Krankheiten, 1911, 6, 249-326
- ROTTERS, G.: Die Ergebnisse einer Querschnittuntersuchung an zehn- bis elfjährigen Kindern mit den "Tests of Primary Mental Abilities" von L.L. Thurstone und Th. G. Thurstone. Münster 1966, unveröffentlicht
- ROTH, E./OSWALD, W.D./DAUMENLANG, K.: Intelligenz. Stuttgart 1972
- RUDINGER, G.: Tendenzen und Entwicklungen entwicklungspsychologischer Versuchsplanung - Sequenzanalysen. Psychologische Rundschau, 1981, 32, 118-136
- SADER, M.: Möglichkeiten und Grenzen psychologischer Testverfahren. Bern 1961
- SCHAIK, K.W.: Can the Longitudinal Method Be Applied to Psychological Studies of Human Development? in MÜNKES, F.J./HARTUP, W.W./WIT de, J. (Eds): Determinants of Behavioral Development. New York 1972, 3-22
- SCHAIK, K.W.: Methodische Probleme bei der deskriptiven entwicklungspsychologischen Untersuchung des Erwachsenen- und Greisenalters. in BALTERS, P.B./ECKENBERGER, L.H. (Hrsg.): Entwicklungspsychologie der Lebensspanne. Stuttgart 1979, 179-208
- SCHMID, F.W.: Intelligenzforschung. Zürich/Köln 1977
- SCHRÖTER, G.: Zensurengebung - Allgemeine und fachspezifische Probleme. Kastellaun 1977
- SELG, H./BAUER, W.: Forschungsmethoden der Psychologie. Stuttgart 1971

- SHIELDS, J.: Monozygotic Twins Brought Up Apart and Brought Up Together. London 1962
- SIEGEL, S.: Nonparametric Statistics for the behavioral Sciences. Tokyo 1956
- SIMONS, D./LANGHEINRICH, D.: Fehleranalyse bei den Analyseaufgaben des Intelligenz-Struktur-Test (I-S-T-70) von Amthauer. in THOMAS, A./BRACKHANE, R.: Wahrnehmen - Urteilen - Handeln. Bern 1980, S. 179-210
- SKARABIS, H.: Mathematische Grundlagen und praktische Aspekte der Diskrimination und Klassifikation. Würzburg 1970
- SKEELS, H.M.: Some Iowa Studies of the Mental Growth of Children in Relation to Differentials of the Environment: A Summary. Yearb. nat. Soc. Stud. Educ., 1940, 39, 281-308
- SKOWRONEK, H. (Hrsg.): Umwelt und Begabung. Stuttgart 1973
- SMITHERS, A.G./COOPER, H.J.: Personality and Season of Birth. Journal of Social Psychology, 1978, 105, 237-241
- SPLITZ, R.A.: Vom Säugling zum Kleinkind. Naturgeschichte der Mutter-Kind-Beziehungen im ersten Lebensjahr. Stuttgart ³1972
- SPLITZNAGEL, A.: Grundlagen, Ergebnisse und Probleme der Formdeutungsverfahren. in HEISS, R. (Hrsg.): Handbuch der Psychologie, Bd. 6, Göttingen 1964, 556-618
- STAKE, R.E.: Testing Hazards in Performance Contracting. in PAYNE, D.A./McMORRIS, R.F. (Eds.): Educational and Psychological Measurement - Contributions to Theory and Practice. New York ²1975, S. 279-291
- STERN, W.: Zur Theorie der Intelligenz-Konstanz. Zeitschrift für angewandte Psychologie und psychologische Sammelforschung, 1925, 34, 146-155

- STERNBERG, R.J.: Editor's Note. Intelligence, 1981, 5, 1
- STOLL, F.: Testgebrauch in der schweizerischen Beratungspraxis. in PULVER, U./LANG, A./SCHMID, F.W. (Hrsg.): Ist Psychodiagnostik verwertbar? Bern 1978, S. 321-339
- STROTMANN, M.: Leistungs- und Sozialverhalten von Schulneulingen im Klassenverband. Münster 1960, unveröffentlicht
- STURZENBECHER, K.: Intelligenzhöhe, Begabungsrichtungen und Erfolg von Kindern im 9. Besuchsjahr allgemeinbildender Schulen. Münster 1968, unveröffentlicht
- TENT, L.: Die Auslese von Schülern für weiterführende Schulen - Möglichkeiten und Grenzen. Göttingen ²1969
- THOMAE, H.: 50 Jahre Längsschnittforschung: Ein Beitrag zur Trendanalyse der Entwicklungspsychologie. in MONTADA, L. (Hrsg.): Brennpunkte der Entwicklungspsychologie. Stuttgart 1979, 31-41
- THOMPSON, D.: Season of Birth and Success in Secondary School. Educational Research 1971, 14, 56-60
- THORNDIKE, R.L.: Intellectual Status and Intellectual Growth. Journal of Educational Psychology, 1966, 57, 121-127
- THURSTONE, L.L.: Primary Mental Abilities. Chicago 1938
- THURSTONE, L.L./THURSTONE, T.G.: Primary Mental Abilities for Ages 7 to 11. Chicago 1954
- THURSTONE, T.G.: PMA Primary Mental Abilities. Technical Report. Chicago, Illinois 1965
- TIETZE, W.: Chancenungleichheit bei Schulbeginn. Düsseldorf 1973

- TURNBULL, W.W.: Intelligence Testing in the Year
2000. Intelligence, 1979, 3, 275-282
- TYLER, L.E.: The Psychology of Human Differences.
New York ³1965
- WALTER, M.: Auslesetest für Schulneulinge - Form B.
Testreferat der Stadt München o.J.
- WECHSLER, D.: Die Messung der Intelligenz Erwachsener
(HAWIE). Bern ³1964 (Handbuch)
- WECHSLER, D.: WAIS Wechsler Adult Intelligence Scale.
New York 1955
- WECHSLER, D.: WISC-R Wechsler Intelligence Scale for
Children - Revised. New York 1984
- WEDGE, P./PROSSER, H.: Born to Fail? London 1973
- WEINERT, F.: Einführung in das Problemgebiet der
pädagogischen Psychologie des schulischen
Lernens. in WEINERT, F. (Hrsg.): Pädagogische
Psychologie, Köln ⁴1969, 13-41
- WEINERT, L.: Die Konstanz des Intelligenzquotienten
nach neueren amerikanischen Untersuchungen.
Zeitschrift für angewandte Psychologie und
psychologische Sammelforschung, 1925, 34, 156-169
- WENDT, H.W.: Season of Birth, Introversion, and
Astrology: A Chronobiological Alternative.
Journal of Social Psychology, 1978, 105, 243-247
- WENDT, H.W.: Early Circannual Rhythms and Adult
Human Behaviour: Components of a Chronobe-
havioural Theory, and Critique of Persistent
Artifacts. International Journal of Chrono-
biology, 1974, 2, 57-86
- WILLERMAN, L.: The Psychology of Individual and
Group Differences. San Francisco 1979
- WINDHEUSER, H.-J.: Erlebte Selbstverantwortlichkeit
für Erfolge und Mißerfolge in Leistungs-
situationen - im Zusammenhang mit Reife des mo-
ralischen Urteils, Intelligenz, Schulerfolg,
Beruf des Vaters und Konfession. Münster 1968,
unveröffentlicht

- WITTMANN, B.: Vom Sinn und Unsinn der Hausaufgaben. Neuwied 1964
- ZELNIK, M.: Socio-Economic and Seasonal Variations in Births: A Replication. Milbank Memorial Fund Quarterly 1969, 47, 159-165
- ZIMMER, D.E.: Der Streit und die Intelligenz. IQ: ererbt oder erworben? München 1975
- ZINGELER, U.: Erfolg und Versagen von regulär eingeschulden und zurückgestellten Kindern in der Unterstufe der Volksschule. Psychol. Inst. d. Univ. Münster, 1965, unveröffentlicht
- ÜBERLA, K.: Faktorenanalyse. Berlin/Heidelberg ²1971

A N H A N G

Anlage 1

Beruf des Vaters (Codierung, Version I2)
Frage 29 des Interessen-Milieu-Fragebogens (IMF)

- 11 - Höhere Beamte
- 12 - leitende Angestellte
- 16 - selbständige Akademiker
- 21 - mittlere und gehobene Akademiker
- 22 - gehobene Selbständige
- 26 - mittlere Selbständige
- 31 - untere Beamte mit Schreibtisch-
tätigkeit
- 32 - mittlere kaufm. und techn. Ange-
stellte
- 33 - untere kaufm. und techn. Ange-
stellte
- 34 - Meister und solche, die auf un-
terer Ebene eine Führungsposi-
tion innehaben
- 41 - untere Beamte mit manueller Tä-
tigkeit
- 43 - untere kaufm. und techn. Angestellte
mit manueller Tätigkeit
unterste ausführende Angestellte
- 45 - Facharbeiter oder Handwerker
- 46 - kleinere Selbständige
- 55 - ungelernte und angelernte Arbeiter
- 00 - keine Information

Nur Code 1. Spalte

- 1 - Berufe mit Hochschulabschluß oder
Studium
- 2 - Führungspositionen auf mittlerer
Ebene
- 3 - "Stehkragenberufe"
- 4 - manuelle Berufe
- 5 - ausführende Tätigkeit ohne Berufs-
ausbildung
- 0 - keine Information

ANLAGE 2

Diskriminanzanalyse nach

IBM Technical Newsletter vom 14.2.1969

Discriminant Analysis

In the Scientific Subroutine Package, discriminant analysis is normally performed by calling the following three subroutines in sequence:

1. DMATX - to compute means of variables in each group and a pooled dispersion matrix
2. MINV - to invert the pooled dispersion matrix
3. DISCR - to compute coefficients of discriminant functions and evaluate the functions for each observation (individual)

Subroutine DMATX

This subroutine calculates means of variables in each group and a pooled dispersion matrix for the set of groups in a discriminant analysis.

For each group $k = 1, 2, \dots, g$, the subroutine calculates means and sums of cross-products of deviations from means as shown below.

Means:

$$\bar{x}_{jk} = \frac{\sum_{i=1}^{n_k} x_{ijk}}{n_k} \quad (1)$$

where n_k = sample size in the k^{th} group
 $j = 1, 2, \dots, m$ are variables

Sum of cross-products of deviations from means:

$$S_k = \left\{ s_{j\ell} \right\} = \sum_i (x_{ijk} - \bar{x}_{jk})(x_{i\ell k} - \bar{x}_{\ell k}) \quad (2)$$

where $j = 1, 2, \dots, m$
 $\ell = 1, 2, \dots, m$

The pooled dispersion matrix is calculated as follows.

$$D = \frac{\sum_{k=1}^g S_k}{\sum_{k=1}^g n_k - g} \quad (3)$$

where g = number of groups

Subroutine DISCR

This subroutine performs a discriminant analysis by calculating a set of linear functions that serve as indices for classifying an individual into one of K groups.

For all groups combined, the following are obtained.

Common means:

$$\bar{x}_j = \frac{\sum_{k=1}^g n_k \bar{x}_{jk}}{\sum_{k=1}^g n_k} \quad (1)$$

where g = number of groups

$j = 1, 2, \dots, m$ are variables

n_k = sample size in the k^{th} group

$\bar{x}_{jk}$ = mean of j^{th} variable in k^{th} group

Generalized Mahalanobis D^2 statistics, V :

$$V = \sum_{i=1}^m \sum_{j=1}^m d_{ij} \sum_{k=1}^g n_k (\bar{x}_{ik} - \bar{x}_i)(\bar{x}_{jk} - \bar{x}_j) \quad (2)$$

where d_{ij} = the inverse element of the pooled dispersion matrix D

V can be used as chi-square (under assumption of normality) with $m(g-1)$ degrees of freedom, to test the hypothesis that the mean values are the same in all the g groups for these m variables. For each discriminant function $k = 1, 2, \dots, g$, the following statistics are calculated.

Coefficients:

$$C_{ik\alpha} = \sum_{j=1}^m d_{ij} \bar{x}_{jk} \quad (3)$$

where $i = 1, 2, \dots, m$
 $k = k_\alpha$

Constant:

$$C_{ok\alpha} = -1/2 \sum_{j=1}^m \sum_{l=1}^m d_{jl} \bar{x}_{jk} \bar{x}_{lk} \quad (4)$$

For each i^{th} case in each k^{th} group, the following calculations are performed.

Discriminant functions:

$$f_{k\alpha} = \sum_{j=1}^m C_{jk\alpha} x_{ijk} + C_{ok\alpha} \quad (5)$$

where $k_\alpha = 1, 2, \dots, g$

Probability associated with largest discriminant function:

$$P_L = \frac{1}{\sum_{k=1}^g e^{(f_{k\alpha} - f_L)}} \quad (6)$$

where f_L = the value of the largest discriminant function

L = the subscript of the largest discriminant function

Lebenslauf

Am 18.11.1935 wurde ich als erstes von vier Kindern des 1947 in Kriegsgefangenschaft verstorbenen Drehers Friedrich Niemeier und seiner Ehefrau Maria geb. Recker in Enniger (Westf.) geboren. Von 1943 - 1951 besuchte ich die Volksschule in Oelde (Westf.). Anschließend begann ich eine Lehre als technischer Zeichner, die ich 1954 mit der Gehilfenprüfung abgeschlossen habe. Von 1954 - 1965 arbeitete ich als technischer Zeichner bzw. Konstrukteur in diesem Beruf. Neben der Berufstätigkeit besuchte ich ab 1959 das Abendgymnasium der Stadt Duisburg, wo ich 1964 die Reifeprüfung bestand.

Im Sommersemester 1965 begann ich an der Westfälischen - Wilhelms - Universität in Münster mit dem Studium der Psychologie. Ende des 6. Semesters, Wintersemester 1967/68, bestand ich dort die Diplom-Vorprüfung und nach dem 9. Semester, Wintersemester 1969/70, die Diplom-Hauptprüfung für Psychologen. Nach dem Examen habe ich mein Studium bis zum Ende des Wintersemesters 1976/77 fortgesetzt.

Seit meinem 2. Semester war ich studentische Hilfskraft am Psychologischen Institut der Universität Münster. Dort habe ich - nach meinem Hauptexamen als wissenschaftlicher Mitarbeiter - bis 1972 an einer Längsschnittstudie zur Bedingungsanalyse von Schulerfolg und Schulversagen unter der Leitung von Frau Prof. Dr. Lilly Kemmler mitgearbeitet. Von 1973 - 1975 war ich als wissenschaftlicher Mitarbeiter an dem Projekt "Wissenschaftliche Begleitung Kollegstufe NW", Leitung Prof. Dr. Blankertz, tätig.

Vom Sommersemester 1977 bis Wintersemester 1980/81 war ich als wissenschaftliche Hilfskraft bzw. Verwalter einer wissenschaftlichen Assistentenstelle an der Universität-Gesamthochschule-Paderborn beschäftigt. Seit dem Sommersemester 1981 führe ich dort Lehraufträge im Fach Psychologie aus.

Seit 1975 bin ich verheiratet.

FOTODRUCK J. MAINZ – Telefon: (02 41) 2 73 05
Neupforte 13, Nähe Markt, 5100 Aachen